UNIVERSIDAD CARLOS III DE MADRID
Departamento de Teoría de la Señal y Comunicaciones

DOCTORAL THESIS

BAYESIAN NONPARAMETRIC MODELING OF PSYCHIATRIC DISORDERS

Author: MARÍA ISABEL VALERA MARTÍNEZ
Supervised by: FERNANDO PÉREZ CRUZ
NOVEMBER 2014

Tesis Doctoral:	BAYESIAN NONPARAMETRIC MODELING OF PSYCHIATRIC DISORDERS
Autor:	María Isabel Valera Martínez
Director:	D. Fernando Pérez Cruz
Fecha:	

Tribunal

Presidente:

Vocal:

Secretario:

Acknowledgements

First of all, I would like to express my deep gratitude to the two people that have made this thesis possible: my PhD supervisor, Fernando Pérez Cruz; and my college Francisco J. R. Ruiz. I would like to thank Fernando for his guidance, unconditional support and, overall, for his confidence in me. I would also like to thank Francisco for his patience and enthusiasm that have helped me to overcome this process.

I would also like to thank everyone in our research group for making the work hours much more pleasant. Antonio and Joaquín, thank you for giving me the chance to join this group and helping me to grow up as a researcher. For the rest of the members (all, current and past) of this group, thanks for making me laugh even in the days I thought I could not. I take with me a lot of moments and good friendships.

I would also like to thank all these people that accompany me wherever I go, these people that take care of me everyday: Rafa, my family, Marta, Mari, Fagi, Wilton (also called "mi bichico"), Grace, etc. I would like to specially thank my family to be always there to support and encourage me. Finally, I would like to thank Rafa (and our *little* Otto and Dona) for... too many reasons to be told.

Abstract

Mental health care has become one of the major priorities in developed countries, where the annual budgets assigned to mental health care reach hundreds of billion of dollars. Due to lack of laboratory tests as objective diagnostic criteria, there is not consensus among the psychiatrists either on the diagnostic criteria or the treatments. As a consequence, there exists an increasing interest in improving both the detection and treatment of mental disorders. This thesis is an interdisciplinary work, in which we study the causes behind suicide attempts and provide thorough analysis of pathological and comorbidity patterns of mental disorders. The final goal of this study is to help psychiatrists detect people with higher risk and guide them to improve treatments. To this end, we apply latent feature modeling to the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC), which collects information about the mental health of the U.S. population. In order to avoid the model selection step needed to infer the number of variables in the latent feature model, we make use of the Indian Buffet Process (IBP) [27]. However, the discrete nature of the database does not allow us to use the standard Gaussian observation model, and therefore, we need to adapt the observation model to discrete random variables.

In a first step, we propose an IBP model for categorical observations, which are the most common in the NESARC. We consider two likelihood observation models: a multinomial-logit and a multinomial-probit model. We derive efficient Monte-Carlo Markov chain (MCMC) inference algorithms that resort to either the Laplace approximation or the expectation propagation (EP) algorithm to compute the marginal likelihood. We also derive a variational inference algorithm that provides a less expensive, in terms of computational complexity, alternative to the samplers. Afterwards, to account for all the available information about the subjects (that includes also non categorical observations, such as age, incomes or education level), we extend the IBP observation model to handle mixed continuous (real-valued and positive real-valued) and discrete (categorical, ordinal and count) observations. This model keeps the properties of conjugate models and allows us to derive an inference algorithm that scales linearly with the number of observations. Finally, we present the experimental results obtained after applying the proposed models to the NESARC database, studying both the hidden causes behind suicide attempts and the pathological and comobidity patterns of mental disorders.

Resumen

La salud mental se ha convertido en una de las principales prioridades de los países desarrollados, los cuales dedican anualmente cientos de miles de millones de dólares al cuidado de la misma. Debido a la falta de pruebas de laboratorio como criterios objetivos para el diágnostico de los desórdenes mentales, existe una falta de consenso tanto en los criterios de diagnóstico como en los tratamiento. Esta tesis es un trabajo interdisciplinario que tiene como próposito encontrar las causas latentes detrás de los intentos de suicidio y proveer de un profundo análisis sobre los patrones, tanto patológicos como de comorbidad, de los desórdenes psiquiátricos. Como objetivo final de este trabajo, prentendemos ayudar a los psiquiátras a detectar aquellas personas con mayor riesgo de sufrir de desórdenes mentales, y guiarlos en la cartegorización y los tratamientos para dichos desórdenes. Para ello, aplicamos modelado de características latentes a la base de datos NESARC (National Epidemiologic Survey on Alcohol and Related Conditions), la cual contiene información sobre la salud mental de una muestra representativa de la población estadounidense. Con el fin de evitar fijar la complejidad del modelo *a priori*, recurrimos al *Indian Buffet Process* (IBP) [27]. Sin embargo, debido a la naturaleza discreta de la base de datos, debemos adaptar a observaciones discretas el modelo de observación del IBP, que normalmente asume verosmilitudes Gaussianas.

Inicialmente, adaptamos el modelo de observación del IBP a datos categóricos, los más comunes en la NESARC. Para ello, consideramos dos funciones de verosimilitud (la *multinomial-logit* y la *multinomial-probit*) y desarrollamos algoritmos de inferencia basados en muestreo (*Monte-Carlo Markov chain*) los cuales recurren a la aproximación de Laplace o al algoritmo *Expectation Propagation* para calcular la verosimilutud marginal. Adicionalmente, derivamos un algoritmo variacional que presenta menor complejidad que los algoritmos de muestreo. Después, con el fin de tener en cuenta en nuestro análisis toda la información disponible en la base de datos (que incluye otras variables no categóricas como la edad, los ingresos anuales o el nivel de estudios), proponemos un modelo de observación para el IBP que permite manejar bases de datos heterogéneas. Este modelo mantiene las propiedades de los modelos conjugados y permite derivar un algoritmo de inferencia de complejidad lineal con el número de observaciones. Finalmente, analizamos los resultados obtenidos al aplicar los modelos propuestos a la base de datos NESARC, estudiando tanto las causas latentes del suicidio como los patrones patológicos y de comorbidad de los desórdenes mentales.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Motivation

Psychiatric disorders, characterized by sustained abnormal changes in mood, thinking or behavior, contribute to disability in developed countries [5]. As an example, approximately one out of four U.S. adults reported suffering from mental disorders in 2004 [5], and according to the U.S. Agency for Healthcare Research and Quality, the total cost of mental health care in the U.S. in 2006 was $57.5 billion, which is equivalent to the cost of cancer care [78]. In addition to this quantity, we also need to take into account the cost of treatments for mental health and substance abuse which was estimated in 2005 in $135 billion [50]. One might think that the situation of the U.S. is exceptional but a similar obervation can be made about Europe, where the total cost of mental health care reached almost $170 billion in 2005 [77].

Several studies have stated suicide as an outcome of psychiatric disorders, finding that most of psychiatric disorders have an increased risk of suicide [28]. According to the World Health Organization, almost one million people commit suicide every year, which is more than the number of people that die in homicides and war combined. In addition, 10 to 20 million people attempt suicide [2]. As a consequence, attempt suicide prevention is one of the top public health priorities in developed countries. The current strategies for suicide prevention have focused mainly on the treatment of the suicidal behaviors themselves [14], and also on both the detection and treatment of mental disorders [81]. A high proportion of suicide attempters (82%) suffered from comorbid mental disorders [80]. However, despite prevention efforts including improvements in the treatment of depression, the lifetime prevalence of suicide attempts in the U.S., where more than 34,000 suicides

occur and over 370,000 individuals are treated for self-inflicted injuries in emergency rooms every year [1], has remained unchanged over the past decade [38]. This suggests that there is a need to improve the understanding of the risk factors for suicide attempts as well as the psychiatric disorders, particularly in non-clinical population.

Although significant advances in neuroscience and genetics have been made in recent years, psychiatric classification is still nowadays performed according to diagnostic criteria based on clinical consensus. These diagnostic criteria, standardised by the Diagnostic and Statistical Manual of Mental Disorders (DSM), do not totally agree with findings emerging from clinical neuroscience and genetics [32]. As a consequence, laboratory tests are not used as objective diagnostic criteria, which is why the current classification system is subject to the ongoing controversy [59]. Hence, in order to improve the categorization of mental disorders, and also in order to advance in research that connects neuroscience and genetics with mental health care, a better understanding of pathological and comorbid patterns of psychiatric disorders is essential.

Clinical experience and several studies suggest that the analysis of co-occurring or comorbid psychiatric disorders may have etiologic and treatment implications. As a consequence, in 2001/2002, the National Institute on Alcohol Abuse and Alcoholism (NIAAA) conducted the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) with the objective of providing a better understanding of comorbidity in psychiatry. The NESARC collects information about the mental health of the U.S. population through nearly 3,000 questions regarding, among others, their way of life, their medical conditions, depression and other mental disorders. We provide further details on the NESARC database below. The public availability of the NESARC database has led to a battery of works that cover topics such as comorbidity of psychiatric disorders with other drug use disorders, mood and anxiety disorders, and personality disorders. These studies suggest that understanding the underlying interrelationships among psychiatric disorders can be useful to improve the diagnostic classification system and guide treatment approaches for each disorder [7].

Due to the controversy around the diagnostic criteria of mental disorder and the lack of objective diagnostic criteria, statistical analysis of psychiatric data plays an important role in understanding mental disorders, as the long joint history between psychiatry and statistics shows [60]. Hence, initiatives such as NESARC appear as interesting and challenging chances for applying machine learning and data mining techniques, which have

been successfully applied to diverse health care related problems such as hypothesis generation [55], diseases evolution analysis [74], diagnosis [53] or gene expression analysis [69].

In this context, probabilistic modeling appears as a powerful framework for modelling, visualising, and understanding these data sets. Probabilistic modeling has been applied in many areas of computer science, including machine learning, data mining, natural language processing, computer vision, and image analysis [89] but, to the best of our knowledge, it has not been applied to psychiatric data. Data in the real world almost always involves uncertainty, which may come from noise in the measurements, missing information, or from the fact that we only have a randomly sampled subset from a larger population. Probabilistic models are an effective approach for understanding such data, by incorporating our assumptions and prior knowledge of the world. All these properties make probabilistic modeling an ideal candidate to model and analyse psychiatric databases.

In this thesis, we aim at exploiting the properties of probabilistic modeling to thoroughly analyze the hidden causes behind suicide attempts and the pathological and comorbid patterns of psychiatric disorders. To this end, we apply latent feature modeling to the data collected in the NESARC with the aim of finding the latent or hidden variables that explain the data. These latent variables can be understood as latent properties of the objects being modeled that have not been directly observed, or as hidden causes behind the observed data. In order to avoid the model selection step needed to infer the number of variables in the latent feature model, we make use of Bayesian nonparametric (BNP) tools, which allow an open-ended number of degrees of freedom in a model [34]. Specifically, our starting point is the Indian Buffet Process (IBP) [27], because it allows us to infer which latent features influence the observations and how many features there are. An overview on BNP models is provided in Chapter 2.

1.1.1 NESARC database

The NESARC was thought to determine the magnitude of alcohol use disorders and their associated disabilities in the general population and in subgroups of the population. Two waves of interviews have been conducted for this survey (first wave in 2001-2002 and second wave in 2004-2005). In the current work, we only use the data from the first wave, for which 43,093 people were selected to represent the non-institutionalized U.S. population above 18 years old. This wave of data is currently available at: *http://aspe.hhs.gov/hsp/06/catalog-ai-an-na/nesarc.htm*.

Through 2,991 entries, the NESARC collects data on the background of participants, alcohol and other drug consumption and abuse, medicine use, medical treatment, mental disorders, phobias, family history, etc. In the extensive battery of questions, the survey also includes a question about having attempted suicide as well as other related questions such as 'felt like wanted to die' and 'thought a lot about own death'. It mainly contains yes-or-no questions and some multiple-choice answers. Furthermore, the NESARC contains questions associated to the criteria set forth in the American Psychiatric Associations DSM-IV for the following psychiatric disorders:

- Five substance disorders: alcohol abuse, alcohol dependence, drug abuse, drug dependence and nicotine dependence.
- Five mood disorders: major depressive disorder, bipolar I and bipolar II disorders, dysthymia, and hypomania.
- Four anxiety disorders: panic with and without agoraphobia, social phobia, specific phobia, and generalized anxiety.
- Seven personality disorders: avoidant, dependent, obsessive compulsive, paranoid, schizoid, histrionic, and antisocial disorders.

NESARC's diagnostic classifications were based on the Alcohol Use Disorder and Associated Disability Interview Schedule DSM-IV (AUDADIS-IV), which is a semistructured diagnostic interview schedule designed for use by lay interviewers.

1.2 Organization

This thesis is an interdisciplinary work in which we apply probabilistic modeling to psychiatric data. As a consequence, we have structured this thesis into a machine learning and a psychiatry part. Specifically, the first part comprises the technical details corresponding to the machine learning discipline and consists of Chapters 2 to 6; and the second part corresponds to the psychiatric contributions of the thesis and includes Chapters 7 to 9.

In Chapter 2, we begin with an overview of Bayesian nonparametric tools. Specifically, we revise the basic principles of Bayesian nonparametric models, and review two of the most popular Bayesian nonparametric models: the Dirichlet process (DP) and the Indian Buffet process (IBP).

In Chapter 3, we propose an IBP based model suited for psychiatric data. To this end, we extend the IBP model in three ways. First, we adapt

the observation model to account for categorical observations, which are the most common data in NESARC database. In particular, we consider two likelihood observation models for categorical observations: a multinomial-logit and a multinomial-probit model. Then, we include a bias term in the IBP model, i.e., a latent variable that is active for all the objects, that helps us to model those people that do not suffer from any mental disorder, allowing us to interpret latent variables as latent disorders or hidden causes behind psychiatric disorders. Finally, since the disorders are not thought to be on/off diagnostics but rather manifestations or indicators of underlying continuous variables that represent predispositions to certain types of psychopathology, we adapt the IBP model to allow real-valued latent variables. In our model, if we interpret latent variables as latent disorders, once a subject has a latent variable active, its value indicates the grade of severity. We limit the latent variables to be between 0 and 1, which helps to interpret the latent variable as a belief in the subject suffering.

In Chapter 4, we derive several inference algorithms for the IBP model for categorical observations. First, we derive three Markov Chain Monte Carlo (MCMC) based inference algorithms: two (approximate) collapsed Gibbs samplers adapted for the two considered likelihood functions under the IBP model with binary latent variables, and a Metropolis-Hastings (MH) based algorithm to infer the real-valued latent variables. Since both the multinomial-logit and the multinomial-probit functions lead to nonconjugate likelihood models, we cannot analytically compute the marginal likelihood. Instead, we derive a Laplace approximation and an expectation propagation (EP) algorithm for approximately collapsing some of the latent variables under, respectively, the multinomial-logit and the multinomial-probit observation models. Second, we derive a variational inference algorithm for the IBP model with the multinomial-logit observation model. This algorithm presents lower computational complexity than the Gibbs samplers, and therefore, allows us to deal with a larger number of observations.

In Chapter 5, we extend the IBP observation model to handle mixed continuous and discrete observations in order to account for all the available information about the subjects (that includes also non categorical observations, such as age, incomes or education level). In particular, the proposed model is able to handle mixed real-valued, positive real-valued, categorical, ordinal and count data. The model keeps the properties of conjugate models, allowing us to derive an inference algorithm that scales linearly with the number of observations.

In the second part of the thesis, we present the experimental results

obtained after applying the proposed models to the NESARC database. In Chapter 6, we provide an analysis of the hidden causes behind suicide attempts, showing that the proposed model is able to detect people that have a higher risk of attempting suicide.

In Chapter 7, we provide an exhaustive analysis of the comorbid patterns among the 20 psychiatric disorders in the NESARC database. In this chapter, we further analyze how different aspects of the people social background (such as marital status, incomes, etc.) affect to the manifestation of the different disorders.

In Chapter 8, we focus on seven personality disorders, studying their pathological and comobid patterns. This analysis includes an evaluation of the diagnostic criteria used in the NESARC to diagnose the seven personality disorders.

Finally, in Chapter 9, we provide a summary with the main contributions and results in the thesis, and some future possible research lines.

1.3 Contributions

The main contributions of this thesis are two-fold. On the one hand, we have the technical contribution concerning machine learning techniques. On the other hand, we have the contributions of the thesis to the state-of-the-art in psychiatry. The technical contributions include:

- An IBP based model (and several extensions) suited for categorical data and the corresponding inference algorithm. Specifically, we derive three MCMC based inference algorithms and a variational inference algorithm.

- Extension of the IBP model to account for heterogeneous databases, keeping the properties of conjugate models and allowing for efficient and fast (linear complexity) inference.

Note that, although we only focus on psychiatric data, the proposed models and related inference algorithms, are general enough to be applicable in other frameworks suitable for categorical or heterogeneous databases. For instance, the extension of the IBP model to account for heterogeneous databases has proved to be successful in estimating missing data in several databases [86].

We next discuss the contributions regarding psychiatry. As we shall show in the second part of the thesis, we obtain not only results in agreement with previous studies but also new insights in the suicide risk detection and

comorbidity pattern analysis, that may help psychiatrists detect people with higher risk and guide them to improve treatments. The main contributions to the psychiatric discipline are summarized below:

- We devise a suicide risk detector that does not only find the hidden causes of suicide attempts but also allows us to detect those subject with higher risk of attempting suicide.
- We perform an exhaustive analysis of comorbidity patterns among 20 psychiatric disorders that allows us to detect those subjects with higher level of suffering. This study also includes how different aspects of the social background of the subjects, such as age and gender, show up in the comorbidity patterns of psychiatric disorders.
- We perform an comprehensive study of both pathological and comorbid patterns among seven personality disorders. This study shows how the seven personality disorders are related among each other, and provides a thorough analysis and evaluation of the criteria used in the NESARC to diagnose these disorders.

Chapter 2

Brief Introduction to Bayesian Nonparametrics

Bayesian nonparametric (BNP) models are being developed in the statistics and machine-learning communities [57] to solve problems such as topic modeling [8], image segmentation [79], speaker diarization [22], and gene-expression modeling [40], among others. BNP models are useful to find out the latent causes and structures behind data, and appear as an alternative to model selection, which is one of the main concerns within the machine learning community and is highly related to problems such as overfitting and underfitting [57]. Examples of model selection include selecting the number of clusters in a clustering problem [52], the number of latent variables in a latent feature model [27], the number of hidden states in a hidden Markov model [23] , or the number of levels in a network [9]. In BNPs, the model complexity is allowed to grow with data size. The central idea behind BNPs is the replacement of the classical finite-dimensional prior distribution with a general stochastic process allowing for an open-ended number of degrees of freedom in a model [35].

BNP are generative models that explain the observed data with a potentially infinite number of parameters. For example, the Dirichlet process (DP) [75] is a BNP prior to cluster data in which the number of clusters is potentially unbounded while the Indian buffet process (IBP) is a latent variable model in which the number of latent variables is potentially unbounded [27]. Hierarchical Dirichlet processes (HDPs) allow, for instance, describing infinite dimensional hidden Markov models (HMMs). The inference process in BNPs jointly provides the model complexity, i.e., the number of components (e.g., the number of clusters and the cluster

assignment for each data point), as well as parameters of the components (e.g., cluster properties such as the mean and covariance in Gaussian mixture models).

Although BNP models were proposed in the seventies [21], they have not received full attention until fairly recently because of their high computational complexity. The underlying stochastic process behind a BNP model has an infinite number of dimensions that makes the computation of the posterior distribution generally expensive. Recent years, developing computationally feasible inference algorithms in BNPs has captured the attention of the machine learning community (see, e.g., [18, 62, 58]) due to the extreme increase of available data. Indeed, a full 90% of all the data in the world has been generated over the last two years [70].

We find two main branches in BNP inference: Markov Chain Monte Carlo (MCMC) based approaches (see, e.g., [18, 62, 27, 95, 87, 83] and the references therein) and variational inference (see, e.g., [10, 36, 19, 58] and the references therein). MCMC based algorithms consist of iteratively sampling from (either sequentially or in blocks) the unknown variables, asymptotically getting samples from the true posterior distribution. Among MCMC based algorithms, Gibbs sampling appears as one of the most popular in BNPs due to its simplicity and because, in its collapsed version, it allows integrating out variables to accelerate the convergence of the MCMC [52, 27, 18]. Variational algorithms usually appear as faster methods than MCMC based approaches, because they tackle the inference task as an optimization problem. They approximate the intractable posterior distribution with a tractable variational distribution by introducing additional independence assumptions that ease the update of the variational parameters. These parameters are typically optimized by minimizing the Kullback-Leibler divergence between the true posterior and the variational distribution. However, by searching only within a restricted class of distributions we might lose some of the expressiveness of the model, leading typically to less accurate results than the MCMC methods, which asymptotically sample from the true posterior [94].

2.1 Dirichlet Process

The Dirichlet process (DP) is currently one of the most popular Bayesian nonparametric models. The DP places a distribution over distributions, i.e. each draw from a Dirichlet process is itself a distribution, and is called Dirichlet process because it has Dirichlet distributed finite dimensional

marginal distributions (refer to [21] for a formal definition of the DP). DPs are used in a wide variety of applications of Bayesian analysis in both statistics and machine learning. The simplest and most popular applications include density estimation and clustering via mixture models [47, 20, 63].

We focus on the DP mixture model for clustering. In a clustering problem, given a set of observation, we aim to divide them into disjoint subsets or clusters. Hence, the main assumption in clustering is that each observation $\mathbf{x}_n$ belongs to a single cluster. Here, the nonparametric nature of the Dirichlet process translates to mixture models with a countably infinite number of components (or clusters). More formally, given a set of N observations, $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, the DP mixture model assumes that each observation $\mathbf{x}_n$ is independently drawn from a distribution $F(\theta_n)$, such that

$$\begin{aligned} &\mathbf{x}_n \sim F(\theta_n), \\ &\theta_n \sim G, \\ &G \sim DP(\alpha, G_0), \end{aligned}$$

where $DP(\alpha, G_0)$ states for a DP with concentration parameter α and base measure G_0. We revise below how to construct the function G, i.e., the DP, but for the time being, let us remark that G is discrete and, therefore, different θ_n ($n \in \{1, \ldots, N\}$) can take simultaneously the same value. Hence, the model above can be seen as a mixture model where the observations $\mathbf{x}_n$ with the same parameter θ_n belong to the same cluster. For instance, in the simplest DP Gaussian mixture model in which we are only interested in estimating the means of the clusters (being the covariance matrix, Σ_x, known), the likelihood function $F(\theta_n)$ is assumed to be Gaussian with mean θ_n and covariance matrix Σ_x, where the means θ_n are distributed as G_0 [63]. Hence, if we assume G_0 to a Gaussian distribution, we can exploit the properties of conjugate models to derive fast and efficient inference algorithms. However, the DP prior is general enough to accommodate for any observation model and prior distribution over the parameters of these models (although the inference of such models is another matter).

2.1.1 The Stick-Breaking Construction

The stick-breaking construction of the DP is an equivalent representation of the DP prior, in which draws from a DP are composed of a weighted sum of point masses [76]. Specifically, the stick-breaking construction of the DP

is given by

$$
\begin{aligned}
v_k &\sim \text{Beta}(1, \alpha), \\
\pi_k &= v_k \prod_{l=1}^{k} (1 - v_l), \\
\theta_k^* &\sim G_0, \\
G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}.
\end{aligned}
$$

This construction can be understood with a *stick-breaking process*, in which, starting with a stick of length 1, at each iteration $k = 1, 2, \ldots$, a piece of length π_k is broken off from current length of the stick (refer o Figure 2.1 for a graphical view). Due to its simplicity, the stick-breaking construction of the DP allows for the development of simple inference algorithms [33].

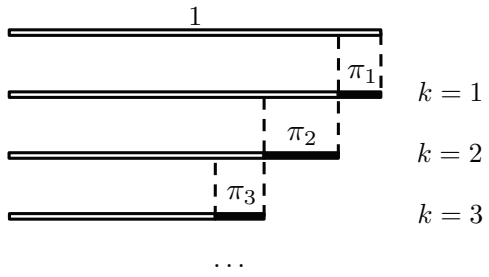


Figure 2.1: Illustration of the stick-breaking construction of the DP.

2.1.2 The CRP and Inference

There are several MCMC based algorithms to perform inference under a DP mixture model, being Gibbs sampling approaches the most popular [33, 52]. We summarize here, one of the simplest Gibbs sampling schemes for inference in DP mixture models. For a better understanding of the algorithm, we introduce the Chinese restaurant process (CRP), which describes the marginal probabilities of the DP in terms of a random partition obtained from a sequence of customers sitting at tables in a restaurant [6]. The CRP allows us to generate samples from a DP in a simple and direct manner.

The CRP receives its name due to a culinary metaphor, in which we have a Chinese restaurant with an infinite number of tables, each of which can

allocate an infinite number of customers (see Figure 2.2). In this metaphor, the first customer enters the restaurant and sits at the first table. The second customer enters and decides either to sit with the first customer, or at a new table. In general, the n-th customer either joins an already occupied table k with probability proportional to the number of customers already sitting there n_k, or sits at a new table with probability proportional to α. This process defines a distribution on partitions and is analogous to the stick-breaking construction of the DP detailed above.

Based on the CRP idea, we can perform inference in an infinite mixture model by iteratively sampling as follows: For $n = 1, \ldots, N$, we assign data point n to an existing cluster or table k with probability proportional to $\frac{n_k}{\alpha+n-1}$ (being n_k the number of customers in table k), or we assign n to a new cluster with probability proportional to $\frac{\alpha}{\alpha+n-1}$. For further details on Gibbs sampling schemes in DP mixture models refer to [33, 52]. Alternatively, a variational inference scheme for the DP mixture model can be found in [10].

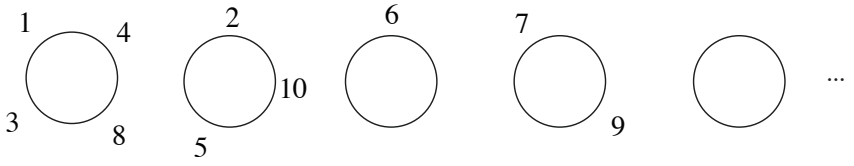


Figure 2.2: A partition induced by the Chinese restaurant process. Numbers indicate customers (objects), circles indicate tables (clusters).

Figure 2.2: A partition induced by the CRP. Numbers indicate customers (objects), circles indicate tables (clusters).

it is identical to the extended Polya urn scheme introduced by Blackwell and MacQueen 1973). Imagine a restaurant with an infinite number of tables, each with an infinite number of seats.[2] The customers enter the restaurant one after another, and each choose a table at random. In the CRP with parameter α, each customer chooses an occupied table with probability proportional to the number of occupants, and chooses the next vacant table with probability proportional to α. For example, Figure 2 shows the state of a restaurant after 10 customers have chosen tables using this procedure. The first customer chooses the first table with probability $\frac{\alpha}{\alpha} = 1$. The second customer chooses the first table with probability $\frac{1}{1+\alpha}$, and the second table with probability $\frac{\alpha}{1+\alpha}$. After the second customer chooses the second table, the third customer chooses the first table with probability $\frac{1}{2+\alpha}$, the second table with probability $\frac{1}{2+\alpha}$, and the new table with probability $\frac{\alpha}{2+\alpha}$. The seating process continues until all customers have seats, defining a distribution over allocations of people to tables and, more generally, objects to classes. Extensions of the CRP and connections to other stochastic processes are pursued in depth by Pitman (2002).

The distribution over partitions induced by the CRP is the same as that given in Equation 5. If we assume an ordering on our N objects, then we can assign them to classes sequentially using the method specified by the CRP, letting objects play the role of customers and classes play the role of tables. The ith object would be assigned to the kth class with probability

$$P(c_i = k | c_1, \ldots, c_{i-1}) = \begin{cases} \frac{m_k}{i-1+\alpha} & k \leq K_+ \\ \frac{\alpha}{i-1+\alpha} & k = K_+ + 1 \end{cases}$$

where m_k is the number of objects currently assigned to class k, and K_+ is the number of classes for which $m_k > 0$. If all N objects are assigned to classes via this process, the probability of a partition of objects $\mathbf{c}$ is that given in Equation 5. The CRP thus provides an intuitive means of specifying a prior for infinite mixture models, as well as revealing that there is a simple sequential process by which exchangeable class assignments can be generated.

2.4 Inference by Gibbs Sampling

Inference in an infinite mixture model is only slightly more complicated than inference in a mixture model with a finite, fixed number of classes. The standard algorithm used for inference in infinite mixture models is Gibbs sampling (Bush and MacEachern, 1996; Neal, 2000). Gibbs sampling

2. Pitman and Dubins, both statisticians at the University of California, Berkeley, were inspired by the apparently infinite capacity of Chinese restaurants in San Francisco when they named the process.

2.2 The Indian Buffet Process

Unsupervised learning aims to recover the latent structure responsible for generating the observed properties of a set of objects. In latent feature modeling, the properties of each object can be represented by an unobservable vector of latent features, and the observed properties are generated from a distribution determined by those latent feature values. Typically, we have access to the set of observations and the main goal of latent feature modeling is to find out the latent variables that represent the data.

The most common nonparametric tool for latent feature modeling is the Indian Buffet Process (IBP). The IBP places a prior distribution over binary matrices, in which the number of rows is finite but the number of columns (features) K is potentially unbounded, that is $K \to \infty$. This distribution is invariant to the ordering of the features and can be derived by taking the

limit of a properly defined distribution over $N \times K$ binary matrices as K tends to infinity [27], similarly to the derivation of the Chinese restaurant process as the limit of a Dirichlet-multinomial model [4]. However, given a finite number of data points N, it ensures that the number of non-zero columns, namely, K_+, is finite with probability one.

Let $\mathbf{Z}$ be a random $N \times K$ binary matrix distributed following an IBP, i.e., $\mathbf{Z} \sim \text{IBP}(\alpha)$, where α is the concentration parameter of the process, which controls the number of non-zero columns K_+. The n-th row of $\mathbf{Z}$, denoted by $\mathbf{z}_n$, represents the vector of latent features of the n-th data point, and every entry nk is denoted by z_{nk}. Note that each element $z_{nk} \in \{0, 1\}$ indicates whether the k-th feature contributes to the n-th data point. Since only the K_+ non-zero columns of $\mathbf{Z}$ contain the features of interest, and due to the exchangeability property of the features under the IBP prior, they are usually grouped in the left hand side of the matrix, as illustrated in Figure 2.3.

$$\mathbf{Z} = \underbrace{\left[\begin{array}{cccc|ccc} z_{11} & z_{12} & \cdots & z_{1K_+} & 0 & 0 & \cdots \\ z_{21} & z_{22} & \cdots & z_{2K_+} & 0 & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \\ z_{N1} & z_{N2} & \cdots & z_{NK_+} & 0 & 0 & \cdots \end{array}\right]}_{K \text{ columns (features)}} \Bigg\} N \text{ data points}$$

K_+ non-zero columns

Figure 2.3: Illustration of an IBP matrix.

2.2.1 The Stick-Breaking Construction

The stick-breaking construction of the IBP is an equivalent representation of the IBP prior, useful for inference algorithms other than Gibbs sampling, such as slice sampling or variational inference algorithms [82, 19].

In this representation, the probability of each latent feature being active is represented explicitly by a random variable. In particular, the probability

of feature z_{nk} taking value 1 is denoted by ω_k, that is,

$$z_{nk} \sim \text{Bernouilli}(\omega_k).$$

Since this probability does not depend on n, the stick-breaking representation explicitly shows that the ordering of the data does not affect the distribution.

The probabilities ω_k are, in turn, generated by first drawing a sequence of independent random variables $v_1, v_2, \ldots$ from a beta distribution of the form

$$v_k \sim \text{Beta}(\alpha, 1).$$

Given the sequence of variables $v_1, v_2, \ldots$, the probability ω_1 is assigned to v_1, and each subsequent ω_k is obtained as

$$\omega_k = \prod_{i=1}^{k} v_i,$$

resulting in a decreasing sequence of probabilities ω_k. Specifically, the expected probability of feature z_{nk} being active decreases exponentially with the index k. The graphical model corresponding to the stick-breaking construction of the IBP is shown in Figure 2.4.

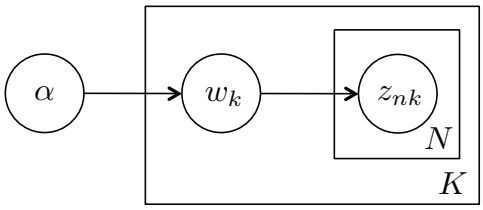


Figure 2.4: Graphical model of the stick-breaking construction of the IBP.

This construction can be understood with the stick-breaking process illustrated in Figure 2.5. Starting with a stick of length 1, at each iteration $k = 1, 2, \ldots$, a piece is broken off at a point v_k relative to the current length of the stick. The variable ω_k corresponds to the length of the stick just broken off, and the other piece of the stick is discarded.

2.2.2 Inference

Markov Chain Monte Carlo (MCMC) methods have been broadly applied to infer the latent structure $\mathbf{Z}$ from a given observation matrix $\mathbf{X}$ (see, e.g.,

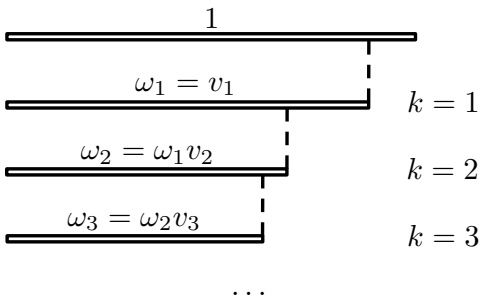


Figure 2.5: Illustration of the stick-breaking construction of the IBP.

in [27, 95, 87, 83]), being Gibbs sampling the standard method of choice. This algorithm iteratively samples the value of each element z_{nk} given the remaining variables, that is, it samples from

$$p(z_{nk} = 1|\mathbf{X}, \mathbf{Z}_{\neg nk}) \propto p(\mathbf{X}|\mathbf{Z})p(z_{nk} = 1|\mathbf{Z}_{\neg nk}), \tag{2.1}$$

where $\mathbf{Z}_{\neg nk}$ denotes all the entries of $\mathbf{Z}$ other than z_{nk}. The conditional distribution $p(z_{nk} = 1|\mathbf{Z}_{\neg nk})$ can be readily derived from the exchangeable IBP and can be written as

$$p(z_{nk} = 1|\mathbf{Z}_{\neg nk}) = \frac{m_{\neg n,k}}{N},$$

where $m_{\neg n,k}$ is the number of data points with feature k, not including n, i.e., $m_{\neg n,k} = \sum_{i \neq n} z_{ik}$. For each data point n, after having sampled all elements z_{nk} for the K_+ non-zero columns in $\mathbf{Z}$, the algorithm samples from a distribution (where the prior is a Poisson distribution with mean α/N) a number of new features necessary to explain that data point.

Although MCMC methods perform exact inference, they typically suffer from high computational complexity. To solve this limitation, variational inference algorithms can be applied instead at a lower computational cost, at the expense of performing approximate inference [36]. A variational inference algorithm for the IBP under the standard Gaussian observation model is presented by [19]. This algorithm makes use of the stick breaking construction of the IBP, summarized above.

Chapter 3

IBP for Categorical Observations

As introduced in Chapter 1, the main goal of this work is to find and interpret the latent patterns behind psychiatric disorders. In this chapter, we propose to model the subjects in the NESARC database using a BNP latent model that allows us to seek hidden causes and compact in a few features the immense redundant information. Our starting point is the IBP [27], because it allows us to infer which latent features influence the observations and how many features there are. As the NESARC database mostly contains yes-or-no questions and some multiple-choice answers, we need an observation model suited for categorical observations. We propose two observation models: a multinomial-logit and a multinomial-probit likelihood model.

Additionally, we extend the IBP model motivated by the specific application of modeling the latent factors behind psychiatric disorders. We extend the IBP model in two ways. First, we add a bias term, which plays the role of a latent variable that is always active. For a discrete observation space, if we do not have a bias term and all latent variables are inactive, the model assumes that all the outcomes are independent and equally likely, which is not a suitable assumption in psychiatry. Second, we consider the latent variables to be bounded real values, instead of on-off latent features. Once a subject activates a latent variable, its value indicates the grade of influence of the latent feature on this subject. Hence, if we interpret a latent feature as a latent disorder, its value indicates the severity of suffering.

3.1 Model Description

Let us assume N objects, where each object is defined by D attributes. We can store the data in an $N \times D$ observation matrix $\mathbf{X}$, in which each D-dimensional row vector is denoted by $\mathbf{x}_n = [x_n^1, \ldots, x_n^D]$ and each entry is denoted by x_n^d. Let us also denote each N-dimensional column vector in $\mathbf{X}$ by $\mathbf{x}^d$. Here, unlike the standard Gaussian observation model, we consider categorical observations, i.e., x_n^d takes values in a finite unordered set $\{1, \ldots, R_d\}$, e.g., $x_n^d \in \{$'blue', 'red', 'black'$\}$. For simplicity and without loss of generality, we assume the same number of categories R in all the dimensions of $\mathbf{X}$, i.e., $R_d = R$. Nevertheless, the following results can be readily extended to a different cardinality per input dimension.

We assume that each observation x_n^d can be explained by a K-length vector of binary latent variables $\mathbf{z}_n = [z_{n1}, \ldots, z_{nK}]$ and the associated factors b_{kr}^d for $r \in \{1, \ldots, R\}$ that weight the contribution of the k-th latent variable to the observation x_n^d. Note that we have a weighting factor b_{kr}^d for each possible value of the observation $x_n^d \in \{1, \ldots, R\}$. We gather the weighting factors associated to the d-th dimension of $\mathbf{X}$, i.e., b_{kr}^d for $d = 1, \ldots, D$, in a $K \times R$ weighting matrix $\mathbf{B}^d$ (being K the number of latent variables). Similarly, we gather the latent binary feature vectors $\mathbf{z}_n$ in an $N \times K$ matrix $\mathbf{Z}$, which follows an IBP with concentration parameter α, i.e., $\mathbf{Z} \sim \text{IBP}(\alpha)$ [27]. We place a Gaussian distribution with zero mean and variance σ_B^2 over the weighting factors b_{kr}^d. The resulting model is shown in Figure 3.1.

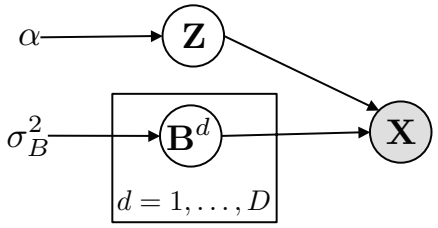


Figure 3.1: Simplest IBP model for Categorical Observation.

For a categorical observation space, the latent model in Figure 3.1 assumes that the observations for an object with no active latent features are independent and equally likely. However, this property does not sound as an appealing outcome when dealing with categorical observations in general, and more so in psychiatry, where only a small fraction of the population suffers from a psychiatric disorder. To solve this limitation, we extend

the model in Figure 3.1 by adding a bias term, which plays the role of a latent feature that is always active and is needed to model the behavior of the objects without any active latent feature. In our application, we make use of the bias term to model the general population that does not suffer from any latent disorder, which allows us to directly interpret the active latent variables as latent features describing disorders. The extended model including the bias term is shown in Figure 3.2 where, similarly to the weighting matrices, we place a Gaussian prior over the bias terms $b_{0r}^d \sim \mathcal{N}(b_{0r}^d|0, \sigma_B^2)$. The bias terms b_{0r}^d are grouped in the K-length vectors $\mathbf{b}_0^d$. Note that we can simplify the notation in Figure 3.2 by assuming an extended latent feature matrix $\mathbf{Z}$ of size $N \times (K+1)$, in which the elements of the first column are equal to one, and D extended weighting matrices $\mathbf{B}^d$ of size $(K+1) \times R$, in which the first row equals the vector $\mathbf{b}_0^d$.

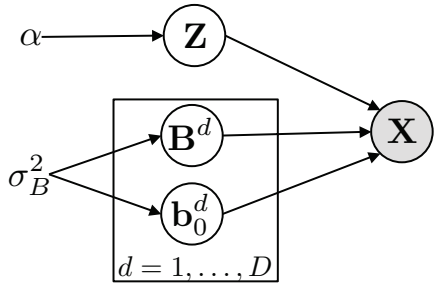


Figure 3.2: IBP model for Categorical observations with bias.

Additionally, we should take into account that in psychiatry the disorders are not thought to be on/off diagnostics, but rather manifestations or indicators of underlying continuous variables that represent predispositions to certain types of psychopathology. Hence, instead of on/off latent features, we extend the IBP model to allow real-valued latent variables. Under this extended model, shown in Figure 3.3, once a subject has a latent variable (or latent disorder) active, its value indicates the severity with what the subject suffers from it. We limit the latent variables to be between 0 and 1, which also helps to interpret the latent variable as a belief in the subject suffering a latent disorder. In particular, we propose an $N \times K$ severity matrix $\mathbf{W}$, where each element $w_{nk} \in [0, 1]$ represents how much the n-th observation is influenced by the k-th latent feature. Similarly to [41], we propose a spike and slab prior for the severity factors to readily account for the subjects that do not suffer from the disorder (spike component), and that allows assigning

a degree of severity for an active latent feature (slab component), i.e.,

$$p(w_{nk}|\gamma_1, \gamma_2, z_{nk}) = (1 - z_{nk})\delta(w_{nk}) + z_{nk}\text{Beta}(w_{nk}|\gamma_1, \gamma_2), \tag{3.1}$$

where $\delta(\cdot)$ is the Kronecker delta function, and γ_1 and γ_2 are hyper-parameters of the model that describe the beta distribution. The combination of the IBP with continuous latent variables has been previously proposed in [41] for a BNP independent component analysis (ICA). In this model, the prior for the latent continuous variables and the IBP matrix are conjugated with a Gaussian likelihood, which significantly differs from our proposal.

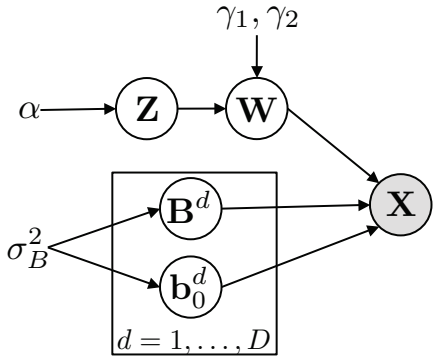


Figure 3.3: Full latent feature model for categorical observations with real-valued latent features.

We assume that the observations x_n^d are independent given the severity matrix $\mathbf{W}$, the weighting matrices $\mathbf{B}^d$ and the vectors $\mathbf{b}_0^d$. Therefore, the likelihood can be factorized as

$$p(\mathbf{X}|\mathbf{W}, \mathbf{B}^1, \ldots, \mathbf{B}^D, \mathbf{b}_0^1, \ldots, \mathbf{b}_0^D) = \prod_{n=1}^{N} \prod_{d=1}^{D} p(x_n^d|\mathbf{w}_n, \mathbf{B}^d, \mathbf{b}_0^d). \tag{3.2}$$

We consider two different likelihood models, a multinomial-logit and a multinomial-probit model. Each of them allows us to derive different inference algorithms, detailed in Chapter 4. There are several alternatives to model categorical observations given the hidden latent features, such as a Dirichlet distribution. However, we prefer the multinomial-logit and the multinomial-probit distributions because, as in the standard Gaussian observation model, the probability distribution of the observations depends on the IBP matrix weighted by some factors, resulting in versatile and flexible likelihood models.

Multinomial-Logit Model. Under the multinomial-logit model the probability of each element x_n^d taking value $r \in \{1, \ldots, R\}$ is given by

$$p(x_n^d = r | \mathbf{w}_n, \mathbf{B}^d, \mathbf{b}_0^d) = \frac{\exp\left(\mathbf{w}_n \mathbf{b}_{\cdot r}^d + b_{0r}^d\right)}{\sum_{r'=1}^{R} \exp\left(\mathbf{w}_n \mathbf{b}_{\cdot r'}^d + b_{0r'}^d\right)}, \tag{3.3}$$

where $\mathbf{b}_{\cdot r}^d$ corresponds to the r-th column vector of $\mathbf{B}^d$, i.e., $\mathbf{b}_{\cdot r}^d = [b_{1r}^d, \ldots, b_{Kr}^d]^\top$. The multinomial-logit model allows the implementation of an efficient Gibbs sampler when the Laplace approximation [48] is used to integrate out the weighting factors and can be efficiently computed using the Matrix Inversion Lemma.

Multinomial-Probit Model. Under the multinomial-probit model, the probability of each element x_n^d taking value $r \in \{1, \ldots, R\}$ can be written as

$$p(x_n^d | \mathbf{w}_n, \mathbf{B}^d, \mathbf{b}_0^d) = \mathbb{E}_{p(u)}\left[\prod_{\substack{r=1 \\ r \neq x_n^d}}^{R} \Phi\Big(u + (b_{0x_n^d}^d - b_{0r}^d) + \mathbf{w}_n(\mathbf{b}_{\cdot x_n^d}^d - \mathbf{b}_{\cdot r}^d)\Big)\right], \tag{3.4}$$

where $\mathbf{w}_n$ stands for the n-th row of matrix $\mathbf{W}$, the auxiliary variable u is distributed as $p(u) = \mathcal{N}(u|0, 1)$, $\mathbb{E}_{p(u)}[\cdot]$ denotes expectation with respect to the distribution $p(u)$, and $\Phi(\cdot)$ denotes the cumulative density function of the standard normal distribution. We use a multivariate cumulative Gaussian likelihood because it is amenable for an EP inference algorithm [26].

Chapter 4

Inference

In this chapter, we derive several algorithms for inferring the latent variables of the models proposed in Chapter 3. First, we focus on the derivation of MCMC based algorithms for the proposed models. Afterwards, we derive a variational inference algorithm for the IBP model fo categorical observations under the multinomial-logit likelihood function.

4.1 MCMC based Inference

We first focus on the simplest model in Figure 3.1, where the unknown variables are the latent matrices $\mathbf{Z}$ and $\{\mathbf{B}^d\}_{d=1}^D$. We remark that the bias term can be directly incorporated into the binary latent matrix $\mathbf{Z}$ and the weighting matrix $\mathbf{B}^d$, such that the extended $\mathbf{Z}$ stands for the $N \times (K+1)$ matrix $[\mathbf{1}\ \mathbf{Z}]$ (being $\mathbf{1}$ a N-length vector whose elements are equal to one), and the extended $\mathbf{B}^d$ stands for the $(K+1) \times D$ matrix $[(\mathbf{b}_0^d)^\top\ (\mathbf{B}^d)^\top]^\top$.

In Section 2.2, we briefly reviewed the collapsed Gibbs sampling algorithm for posterior inference over the latent variables of the IBP. This algorithm samples from

$$p(z_{nk} = 1|\mathbf{X}, \mathbf{Z}_{\neg nk}) \propto p(\mathbf{X}|\mathbf{Z})p(z_{nk} = 1|\mathbf{Z}_{\neg nk}), \tag{4.1}$$

where the marginal likelihood $p(\mathbf{X}|\mathbf{Z})$ is obtained after integrating out the matrices $\mathbf{B}^d$ in

$$p(\mathbf{X}|\mathbf{Z}, \mathbf{B}^1, \ldots, \mathbf{B}^D) = \prod_{n=1}^{N} \prod_{d=1}^{D} p(x_n^d|\mathbf{z}_n, \mathbf{B}^d). \tag{4.2}$$

In the standard Gaussian observation model of the IBP [27], this marginalization can be performed analytically. However, under both the

multinomial-logit and the multinomial-probit models, the marginalization with respect to $\mathbf{B}^d$ is intractable. To solve this limitation, we derive two different algorithms to approximately integrate out the matrices $\mathbf{B}^d$. First, we resort to the Laplace approximation which is suitable for the multinomial-logit model and, then, we derive an Expectation Propagation (EP) algorithm which is suitable for the multinomial-probit model. Note that, although we could also sample from the full join posterior distribution, the high dimensionality of our parameter space causes strong dependences among hyper-parameters and latent variables, resulting in a slow mixing of the chains that requires thousands of posterior draws [64].

The rest of the chapter is organized as follows. First, we provide the details on the derivation of the Laplace and EP approximations in Sections 4.1.1 and 4.1.2, respectively. Afterwards, in Section 4.1.2, we also derive an inference algorithm based on the Metropolis-Hastings (MH) algorithm to jointly sample from the latent features z_{nk} and the severity factors w_{nk} in the full model in Figure 3.3. Finally, in order to evaluate the performance of the Laplace and the EP approximations, we provide a comparison of both approximations on two toy examples in Section 4.1.4.

4.1.1 Laplace Approximation

In this section, we consider the multinomial-logit model in which the probability of each element x_n^d taking value $r \in \{1, \ldots, R\}$ is given by

$$\pi_{nr}^d = p(x_n^d = r|\mathbf{z}_n, \mathbf{B}^d) = \frac{\exp\left(\mathbf{z}_n \mathbf{b}_{\cdot r}^d\right)}{\sum_{r'=1}^{R} \exp\left(\mathbf{z}_n \mathbf{b}_{\cdot r'}^d\right)}. \tag{4.3}$$

Recall that our model assumes independence among the observations given the hidden latent variables. Then, the posterior $p(\mathbf{B}^1, \ldots, \mathbf{B}^D|\mathbf{X}, \mathbf{Z})$ factorizes as

$$p(\mathbf{B}^1, \ldots, \mathbf{B}^D|\mathbf{X}, \mathbf{Z}) = \prod_{d=1}^{D} p(\mathbf{B}^d|\mathbf{x}^d, \mathbf{Z}) = \prod_{d=1}^{D} \frac{p(\mathbf{x}^d|\mathbf{B}^d, \mathbf{Z})p(\mathbf{B}^d)}{p(\mathbf{x}^d|\mathbf{Z})}. \tag{4.4}$$

Hence, we only need to deal with each term $p(\mathbf{B}^d|\mathbf{x}^d, \mathbf{Z})$ individually.

The marginal likelihood $p(\mathbf{x}^d|\mathbf{Z})$, which we are interested in, can be obtained as

$$p(\mathbf{x}^d|\mathbf{Z}) = \int p(\mathbf{x}^d|\mathbf{B}^d, \mathbf{Z})p(\mathbf{B}^d)d\mathbf{B}^d. \tag{4.5}$$

Although the prior $p(\mathbf{B}^d)$ is Gaussian, due to the non-conjugacy with the likelihood term, the computation of this integral, as well as the computation of the posterior $p(\mathbf{B}^d|\mathbf{x}^d, \mathbf{Z})$, turns out to be intractable.

Following a similar procedure as in Gaussian processes for multiclass classification [93], we approximate the posterior $p(\mathbf{B}^d|\mathbf{x}^d, \mathbf{Z})$ as a Gaussian distribution using Laplace's method. In order to obtain the parameters of the Gaussian distribution, we define $f(\mathbf{B}^d)$ as the un-normalized log-posterior of $p(\mathbf{B}^d|\mathbf{x}^d, \mathbf{Z})$, i.e.,

$$f(\mathbf{B}^d) = \log p(\mathbf{x}^d|\mathbf{B}^d, \mathbf{Z}) + \log p(\mathbf{B}^d). \tag{4.6}$$

As proven in Appendix A, the function $f(\mathbf{B}^d)$ is a strictly concave function of $\mathbf{B}^d$ and therefore it has a unique maximum, which is reached at $\mathbf{B}^d_{\text{MAP}}$, denoted by the subscript 'MAP' (*maximum a posteriori*) because it coincides with the mean of the Gaussian distribution in the Laplace approximation. We resort to Newton's method to compute $\mathbf{B}^d_{\text{MAP}}$.

Let us stack the columns of $\mathbf{B}^d$ into $\boldsymbol{\beta}^d$, i.e., $\boldsymbol{\beta}^d = \mathbf{B}^d(:)$ for avid Matlab users. The posterior $p(\mathbf{B}^d|\mathbf{x}^d, \mathbf{Z})$ can be approximated as

$$p(\boldsymbol{\beta}^d|\mathbf{x}^d, \mathbf{Z}) \approx \mathcal{N}\left(\boldsymbol{\beta}^d \,\middle|\, \boldsymbol{\beta}^d_{\text{MAP}}, (-\nabla\nabla f)|_{\boldsymbol{\beta}^d_{\text{MAP}}}\right),$$

where $\nabla\nabla f$ is the Hessian of $f(\boldsymbol{\beta}^d)$. Hence, by taking the second-order Taylor series expansion of $f(\boldsymbol{\beta}^d)$ around its maximum, the computation of the marginal likelihood in (4.5) results in a Gaussian integral, whose solution can be expressed as

$$\begin{aligned}\log p(\mathbf{x}^d|\mathbf{Z}) \approx &-\frac{1}{2\sigma_B^2}\,\text{trace}\left\{(\mathbf{B}^d_{\text{MAP}})^\top \mathbf{B}^d_{\text{MAP}}\right\} + \log p(\mathbf{x}^d|\mathbf{B}^d_{\text{MAP}}, \mathbf{Z}) \\ &- \frac{1}{2}\log\left|\mathbf{I}_{R(K+1)} + \sigma_B^2 \sum_{n=1}^{N}\left(\text{diag}(\widehat{\boldsymbol{\pi}}^d_n) - (\widehat{\boldsymbol{\pi}}^d_n)^\top \widehat{\boldsymbol{\pi}}\right) \otimes (\mathbf{z}_n^\top \mathbf{z}_n)\right|,\end{aligned} \tag{4.7}$$

where $\widehat{\boldsymbol{\pi}}^d_n$ is the vector $\boldsymbol{\pi}^d_n = [\pi^d_{n1}, \pi^d_{n2}, \ldots, \pi^d_{nR}]$ evaluated at $\mathbf{B}^d = \mathbf{B}^d_{\text{MAP}}$, and $\text{diag}(\widehat{\boldsymbol{\pi}}^d_n)$ is a diagonal matrix with the values of $\widehat{\boldsymbol{\pi}}^d_n$ as its diagonal elements. Details of the computation of the Hessian and the gradient of function f are provided in Appendix A.

Similarly as in [27], it is straightforward to prove that the limit of Eq. 4.7 is well-defined if $\mathbf{Z}$ has an unbounded number of columns, that is, as $K \to \infty$. The resulting expression for the marginal likelihood $p(\mathbf{x}^d|\mathbf{Z})$ can be readily obtained from Eq. 4.7 by replacing K by K_+, $\mathbf{Z}$ by the submatrix containing only the non-zero columns of $\mathbf{Z}$, and $\mathbf{B}^d_{\text{MAP}}$ by the submatrix containing the K_+ corresponding rows.

Speeding Up the Matrix Inversion

In this section, we propose a method that reduces the complexity of computing the inverse of the Hessian for Newton's method (as well as its determinant) from $\mathcal{O}(R^3K_+^3 + NR^2K_+^2)$ to $\mathcal{O}(RK_+^3 + NR^2K_+^2)$, effectively accelerating the inference procedure for large values of R.

Let us denote with $\mathbf{Z}$ the matrix that contains only the $K_+ + 1$ non-zero columns of the extended IBP matrix that account for the bias terms. The inverse of the Hessian for Newton's method, as well as its determinant in (4.7), can be efficiently carried out if we rearrange the inverse of $\nabla\nabla f$ as follows:

$$(-\nabla\nabla f)^{-1} = \left(\mathbf{D} - \sum_{n=1}^{N} \mathbf{v}_n \mathbf{v}_n^\top\right)^{-1},$$

where $\mathbf{v}_n = (\boldsymbol{\pi})^\top \otimes \mathbf{z}_n^\top$ and $\mathbf{D}$ is a block-diagonal matrix, in which each diagonal submatrix is given by

$$\mathbf{D}_r = \frac{1}{\sigma_B^2}\mathbf{I}_{K_++1} + \mathbf{Z}^\top \operatorname{diag}(\boldsymbol{\pi}_{\cdot d}^r)\,\mathbf{Z}, \tag{4.8}$$

with $\boldsymbol{\pi}_{\cdot r}^d = \left[\ \pi_{1r}^d, \ldots, \pi_{Nr}^d\ \right]^\top$. Since $\mathbf{v}_n\mathbf{v}_n^\top$ is a rank-one matrix, we can apply the Woodbury identity [97] N times to invert the matrix $-\nabla\nabla f$, similar to the RLS (Recursive Least Squares) updates [30]. At each iteration $n = 1, \ldots, N$, we compute

$$(\mathbf{D}^{(n)})^{-1} = \left(\mathbf{D}^{(n-1)} - \mathbf{v}_n\mathbf{v}_n^\top\right)^{-1} = (\mathbf{D}^{(n-1)})^{-1} + \frac{(\mathbf{D}^{(n-1)})^{-1}\mathbf{v}_n\mathbf{v}_n^\top(\mathbf{D}^{(n-1)})^{-1}}{1 - \mathbf{v}_n^\top(\mathbf{D}^{(n-1)})^{-1}\mathbf{v}_n}. \tag{4.9}$$

For the first iteration, we define $\mathbf{D}^{(0)}$ as the block-diagonal matrix $\mathbf{D}$, whose inverse matrix involves computing the R matrix inversions of size $(K_+ + 1) \times (K_+ + 1)$ of the matrices in (4.8), which can be efficiently solved applying the Matrix Inversion Lemma. After N iterations of (4.9), it turns out that $(-\nabla\nabla f)^{-1} = (\mathbf{D}^{(N)})^{-1}$.

For the determinant in (4.7), similar recursions can be applied using the Matrix Determinant Lemma [29], which states that $|\mathbf{D} + \mathbf{v}\mathbf{u}^\top| = (1 + \mathbf{v}^\top\mathbf{D}\mathbf{u})|\mathbf{D}|$, and $|\mathbf{D}^{(0)}| = \prod_{r=1}^{R} |\mathbf{D}_r|$.

4.1.2 Nested EP

In this section, we adapt the nested EP algorithm introduced in [64] to approximate the marginal likelihood $p(\mathbf{X}|\mathbf{Z})$. To this end, we assume the

multinomial-logit model, being the probability of each observation given by

$$p(x_n^d|\mathbf{z}_n, \mathbf{B}^d) = \mathbb{E}_{p(u)}\left[\prod_{\substack{r=1 \\ r\neq x_n^d}}^{R} \Phi\Big(u + \mathbf{z}_n(\mathbf{b}_{\cdot x_n^d}^d - \mathbf{b}_{\cdot r}^d)\Big)\right], \tag{4.10}$$

where the auxiliary variable u is Gaussian distributed with zero mean and unit variance. Similarly to the multinomial-logit model, the computation of the marginal likelihood, $p(\mathbf{x}^d|\mathbf{Z}) = \int p(\mathbf{B}^d)p(\mathbf{x}^d|\mathbf{Z}, \mathbf{B}^d)d\mathbf{B}^d$, is again intractable, because the prior and likelihood are not conjugate. Instead, we run D parallel nested EP algorithms to compute $p(\mathbf{x}^d|\mathbf{Z})$, being the marginal likelihood $p(\mathbf{X}|\mathbf{Z})$ the product of the individual terms $p(\mathbf{x}^d|\mathbf{Z})$ for $d = 1, \ldots, D$. In the description of the nested EP algorithm, we do not make explicit the dependence on d, unless necessary, to avoid cluttering of notation.

Besides the EP approximation, we could also approximate this posterior using multi-dimensional quadratures [73] or, as before, using the Laplace approximation [26]. We choose the nested EP algorithm, because EP approaches are typically more accurate than the Laplace approximation and computationally less demanding than numerical quadratures [64]. The proposed nested EP consists of two loops, which are described below and summarized in Algorithms 1 and 2. We show in Section 4.1.2 that the complexity of the nested EP is linear in the number of observations. In addition, a heuristic comparison of the performance on two Toy examples of both the Laplace approximation (in Section 4.1.1) and the nested EP approximation is provided in Section 4.1.4. This section also provides a comparison, in terms of flexibility and expressiveness, of the models with both binary latent features and (bounded) real-valued latent variables.

For convenience, we stack the columns of $\mathbf{B}^d$ into the vector $\boldsymbol{\beta}^d$. Note that, given $\mathbf{Z}$, we only need to account for the parameters corresponding to the $K_+ + 1$ active features. To obtain the marginal likelihood, we need to approximate the posterior $p(\boldsymbol{\beta}^d|\mathbf{x}^d, \mathbf{Z})$ with a tractable distribution. The likelihood $p(\mathbf{x}^d|\mathbf{Z}, \boldsymbol{\beta}^d)$ contains a product of non-conjugate terms (sites) [72], denoted by $t_n^d(\boldsymbol{\beta}^d) = p(x_n^d|\mathbf{Z}, \boldsymbol{\beta}^d)$, and hence the posterior can be expressed as

$$p(\boldsymbol{\beta}^d|\mathbf{x}^d, \mathbf{Z}) = \frac{\mathcal{N}(\boldsymbol{\beta}^d|\mathbf{0}, \sigma_B^2\mathbf{I})\prod_{n=1}^{N} t_n^d(\boldsymbol{\beta}^d)}{p(\mathbf{x}^d|\mathbf{Z})}. \tag{4.11}$$

The EP approximation consists on replacing each site $t_n^d(\boldsymbol{\beta}^d)$ with a tractable term $\widetilde{t}_n^d(\boldsymbol{\beta}^d)$, resulting in an approximate distribution that we

denote by $q_{\text{EP}}(\boldsymbol{\beta}^d)$. We choose $\widetilde{t}_n^d(\boldsymbol{\beta}^d)$ to be an unnormalized Gaussian with the $R(K_+ + 1) \times 1$ vector $\boldsymbol{\lambda}_n$ and the $R(K_+ + 1) \times R(K_+ + 1)$ matrix $\widetilde{\boldsymbol{\Pi}}_n$ as natural parameters, and scaling constant $\widetilde{Z}_n$, i.e., $\widetilde{t}_n^d(\boldsymbol{\beta}^d) = \widetilde{Z}_n \mathcal{N}(\boldsymbol{\beta}^d | \widetilde{\boldsymbol{\Pi}}_n^{-1} \widetilde{\boldsymbol{\lambda}}_n, \widetilde{\boldsymbol{\Pi}}_n^{-1})$, yielding

$$\begin{aligned} q_{\text{EP}}(\boldsymbol{\beta}^d) &= \mathcal{N}(\boldsymbol{\beta}^d | \boldsymbol{\Pi}_{\text{EP}}^{-1} \boldsymbol{\lambda}_{\text{EP}}, \boldsymbol{\Pi}_{\text{EP}}^{-1}) \\ &= \frac{1}{Z_{\text{EP}}} \mathcal{N}(\boldsymbol{\beta}^d | \mathbf{0}, \sigma_B^2 \mathbf{I}) \prod_{n=1}^{N} \widetilde{Z}_n \mathcal{N}(\boldsymbol{\beta}^d | \widetilde{\boldsymbol{\Pi}}_n^{-1} \widetilde{\boldsymbol{\lambda}}_n, \widetilde{\boldsymbol{\Pi}}_n^{-1}), \end{aligned} \tag{4.12}$$

where $\boldsymbol{\lambda}_{\text{EP}}$ and $\boldsymbol{\Pi}_{\text{EP}}$ are the natural parameters of the Gaussian distribution $q_{\text{EP}}(\boldsymbol{\beta}^d)$. We choose $\widetilde{Z}_n$ following [71] in order for Z_{EP} to become a good approximation of the marginal likelihood $p(\mathbf{x}^d | \underline{\mathbf{Z}})$.

The EP algorithm chooses the parameters $\widetilde{\boldsymbol{\lambda}}_n$ and $\widetilde{\boldsymbol{\Pi}}_n$ by matching the moments of $p(\boldsymbol{\beta}^d | \mathbf{x}^d, \mathbf{Z})$ and $q_{\text{EP}}(\boldsymbol{\beta}^d)$, which is equivalent to minimizing the Kullback-Leibler divergence $D_{\text{KL}}(p(\boldsymbol{\beta}^d | \mathbf{x}^d, \mathbf{Z}) || q_{\text{EP}}(\boldsymbol{\beta}^d))$. This minimization is solved iteratively for $n = 1, \ldots, N$ [51, 72, 54] (repeating until convergence) as follows:

(i) Define the cavity distribution $q_{\neg n}(\boldsymbol{\beta}^d) \propto q_{\text{EP}}(\boldsymbol{\beta}^d) / \widetilde{t}_n^d(\boldsymbol{\beta}^d)$, in which we have removed one approximate site. The natural parameters of the cavity distribution are $\boldsymbol{\Pi}_{\neg n} = \boldsymbol{\Pi}_{\text{EP}} - \widetilde{\boldsymbol{\Pi}}_n$ and $\boldsymbol{\lambda}_{\neg n} = \boldsymbol{\lambda}_{\text{EP}} - \widetilde{\boldsymbol{\lambda}}_n$.

(ii) Define the tilted distribution $\widehat{p}_n(\boldsymbol{\beta}^d) \propto q_{\neg n}(\boldsymbol{\beta}^d) t_n^d(\boldsymbol{\beta}^d)$ (which includes the true site), and minimize $D_{\text{KL}}(\widehat{p}_n(\boldsymbol{\beta}^d) || q_{\text{EP}}(\boldsymbol{\beta}^d))$ with respect to $q_{\text{EP}}(\boldsymbol{\beta}^d)$.

(iii) Update the approximate site as $\widetilde{t}_n^d(\boldsymbol{\beta}^d) \propto q_{\text{EP}}(\boldsymbol{\beta}^d) / q_{\neg n}(\boldsymbol{\beta}^d)$.

The standard EP algorithm solves Step (ii) by matching the moments between $\widehat{p}_n(\boldsymbol{\beta}^d)$ and $q_{\text{EP}}(\boldsymbol{\beta}^d)$, which is assumed to be tractable. However, in this case, matching these moments is not tractable and we resort to another EP loop, i.e., the inner loop, and hence the name of the algorithm. The inner loop of the nested EP, summarized in Algorithm 2 and detailed in Appendix B, approximates the tilted distribution

$$\widehat{p}_n(\boldsymbol{\beta}^d) = \frac{1}{\widehat{Z}_n} q_{\neg n}(\boldsymbol{\beta}^d) t_n^d(\boldsymbol{\beta}^d) \tag{4.13}$$

by a Gaussian distribution with natural parameters $\widehat{\boldsymbol{\lambda}}_n$ and $\widehat{\boldsymbol{\Pi}}_n$, which is similar to the EP algorithm resulting from a linear binary classifier with a multivariate Gaussian prior and a probit likelihood function in the Gaussian

process setting [61]. Now Step (iii) follows readily, since we can obtain the new natural parameters for the approximate site $\widetilde{t}_n^d(\boldsymbol{\beta}^d)$ as $\widetilde{\boldsymbol{\Pi}}_n^{new} = \widehat{\boldsymbol{\Pi}}_n - \boldsymbol{\Pi}_{\neg n}$ and $\widetilde{\boldsymbol{\lambda}}_n^{new} = \widehat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}_{\neg n}$. A damping factor of η_{O} can be used in this step for numerical stability.

The site parameters $\widetilde{t}_n^d(\boldsymbol{\beta}^d)$ can be updated in parallel for all n, recomputing the parameters of the posterior approximation $q_{\mathrm{EP}}(\boldsymbol{\beta}^d)$ only once per iteration of the outer loop [16, 72]. The approximate posterior parameters are $\boldsymbol{\Pi}_{\mathrm{EP}} = \frac{1}{\sigma_B^2}\mathbf{I} + \sum_{n=1}^N \widetilde{\boldsymbol{\Pi}}_n$ and $\boldsymbol{\lambda}_{\mathrm{EP}} = \sum_{n=1}^N \widetilde{\boldsymbol{\lambda}}_n$. After convergence, the marginal likelihood $p(\mathbf{x}^d|\mathbf{Z})$ can be approximated following [71]. Specifically, the approximate marginal likelihood $p(\mathbf{x}^d|\mathbf{Z})$ can be computed as

$$\begin{aligned}
&\log p(\mathbf{x}^d|\mathbf{Z}) \approx \log Z_{\mathrm{EP}} \\
&= -\frac{1}{2}\log|\boldsymbol{\Pi}_{\mathrm{EP}}| - \frac{KR}{2}\log\sigma_B^2 + \frac{1}{2}\boldsymbol{\lambda}_{\mathrm{EP}}^\top \boldsymbol{\Pi}_{\mathrm{EP}}^{-1}\boldsymbol{\lambda}_{\mathrm{EP}} + \sum_{n=1}^N \log \widetilde{Z}_n,
\end{aligned} \tag{4.14}$$

where

$$\begin{aligned}
\log \widetilde{Z}_n = \log \widehat{Z}_n &+ \frac{1}{2}\boldsymbol{\lambda}_{\neg n}^\top \boldsymbol{\Pi}_{\neg n}^{-1}\boldsymbol{\lambda}_{\neg n} - \frac{1}{2}(\boldsymbol{\lambda}_{\neg n} + \widetilde{\boldsymbol{\lambda}}_n)^\top (\boldsymbol{\Pi}_{\neg n} + \widetilde{\boldsymbol{\Pi}}_n)^{-1}(\boldsymbol{\lambda}_{\neg n} + \widetilde{\boldsymbol{\lambda}}_n) \\
&+ \frac{1}{2}\log|\boldsymbol{\Pi}_{\neg n} + \widetilde{\boldsymbol{\Pi}}_n| - \frac{1}{2}\log|\boldsymbol{\Pi}_{\neg n}|.
\end{aligned} \tag{4.15}$$

Computational complexity

Although the nested EP is similar to the proposed algorithm in [64], the computational complexity is substantially different. The running time of the nested EP for our model is linear in the number of instances (N), while for the Gaussian processes for multiclass classification the computational complexity is cubic. The nested EP for our algorithm needs to integrate out $\boldsymbol{\beta}^d$, which is an $R(K_+ + 1)$-dimensional vector. Note that the outer loop of the proposed nested EP requires one loop in n and, since all the sites $t_n^d(\boldsymbol{\beta}^d)$ are functions of the same $R(K_+ + 1)$-dimensional random vector $\boldsymbol{\beta}^d$, no matrix inversion is needed when we work with the natural parameters of the normal distributions. Each iteration of the inner loop, however, requires the inversion of a matrix of size $R(K_+ + 1) + 1$ (in practice computed using the Cholesky decomposition), which has a complexity of $\mathcal{O}((R(K_+ + 1) + 1)^3)$. The overall complexity of the posterior approximation scales with $DN(R(K_+ + 1) + 1)^3$, because we iterate through the number of samples

Algorithm 1 Outer loop of the nested EP algorithm.

Input: $\mathbf{x}^d$, $\mathbf{Z}$, σ_B^2 (optionally initial site parameters $\widetilde{\mathbf{\Pi}}_n^{ini}$, $\widetilde{\boldsymbol{\lambda}}_n^{ini}$, $\widetilde{\alpha}_{nr}^{ini}$, $\widetilde{\beta}_{nr}^{ini}$)
Output: $p(\mathbf{x}^d|\mathbf{Z})$, $\mathbf{\Pi}_{\text{EP}}$, $\boldsymbol{\lambda}_{\text{EP}}$ (optionally site parameters $\widetilde{\mathbf{\Pi}}_n$, $\widetilde{\boldsymbol{\lambda}}_n$, $\widetilde{\alpha}_{nr}$, $\widetilde{\beta}_{nr}$)
initialize $\widetilde{\mathbf{\Pi}}_n \leftarrow \widetilde{\mathbf{\Pi}}_n^{ini}$, $\widetilde{\boldsymbol{\lambda}}_n \leftarrow \widetilde{\boldsymbol{\lambda}}_n^{ini}$ for $n = 1, \ldots, N$
initialize $\mathbf{\Pi}_{\text{EP}} \leftarrow \frac{1}{\sigma_B^2}\mathbf{I} + \sum_{n=1}^N \widetilde{\mathbf{\Pi}}_n$, $\boldsymbol{\lambda}_{\text{EP}} \leftarrow \sum_{n=1}^N \widetilde{\boldsymbol{\lambda}}_n$
repeat
  for $n = 1, \ldots, N$ (in parallel) **do**
    cavity evaluations: $\mathbf{\Pi}_{\neg n} \leftarrow \mathbf{\Pi}_{\text{EP}} - \widetilde{\mathbf{\Pi}}_n$,
      $\boldsymbol{\lambda}_{\neg n} \leftarrow \boldsymbol{\lambda}_{\text{EP}} - \widetilde{\boldsymbol{\lambda}}_n$
    tilted moments:
      $[\widehat{\mathbf{\Pi}}_n, \widehat{\boldsymbol{\lambda}}_n, \widehat{Z}_n, \{\widetilde{\alpha}_{nr}, \widetilde{\beta}_{nr}\}] \leftarrow$ inner_loop
        $(x_{nd}, \mathbf{w}_n, \mathbf{\Pi}_{\neg n}, \boldsymbol{\lambda}_{\neg n}, \{\widetilde{\alpha}_{nr}^{ini}, \widetilde{\beta}_{nr}^{ini}\})$
    site updates: $\widetilde{\mathbf{\Pi}}_n \leftarrow \eta_{\text{O}}(\widehat{\mathbf{\Pi}}_n - \mathbf{\Pi}_{\neg n}) + (1 - \eta_{\text{O}})\widetilde{\mathbf{\Pi}}_n$,
      $\widetilde{\boldsymbol{\lambda}}_n \leftarrow \eta_{\text{O}}(\widehat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}_{\neg n}) + (1 - \eta_{\text{O}})\widetilde{\boldsymbol{\lambda}}_n$
  end for
  update $\mathbf{\Pi}_{\text{EP}} \leftarrow \frac{1}{\sigma_B^2}\mathbf{I} + \sum_{n=1}^N \widetilde{\mathbf{\Pi}}_n$, $\boldsymbol{\lambda}_{\text{EP}} \leftarrow \sum_{n=1}^N \widetilde{\boldsymbol{\lambda}}_n$
until stopping criterion
for $n = 1, \ldots, N$ (in parallel) **do**
  compute $\log \widetilde{Z}_n$ from (4.15).
end for
compute $\log p(\mathbf{x}^d|\mathbf{Z})$ from (4.14)

N and the dimensionality of the observation vector D. Evaluating the likelihood after convergence of the outer loop requires operations of matrices of size RK_+ within a loop in n, which leads to a complexity scaling with $N(R(K_+{+}1))^3$. Thus, the overall complexity of the full nested EP algorithm to evaluate the marginal likelihood $p(\mathbf{X}|\mathbf{Z})$ is $\mathcal{O}(DN(R(K_+ + 1) + 1)^3)$. The EP procedure can be parallelized in the dimension of the observed instances (D) and in the number of instances N, providing significant savings in runtime complexity.

Furthermore, the site parameters of the inner loop can be stored after each inner EP run and used as starting parameters the next time the inner loop is called [64]. In addition, successive calls to the nested EP algorithm differ in just one element of w_{nk}, which allows reducing the number of outer loop iterations by storing the site parameters $\widetilde{\boldsymbol{\lambda}}_n$ and $\widetilde{\mathbf{\Pi}}_n$ after each nested EP run and continuing from the previous values in the next run. When

Algorithm 2 Inner loop of the nested EP algorithm.

Input: $x_n^d, \mathbf{w}_n, \mathbf{\Pi}_{\neg n}, \boldsymbol{\lambda}_{\neg n}$ (optionally initial site parameters $\widetilde{\alpha}_{nr}^{ini}$, $\widetilde{\beta}_{nr}^{ini}$)
Output: $\widehat{\mathbf{\Pi}}_n, \widehat{\boldsymbol{\lambda}}_n, \widehat{Z}_n$ (optionally site parameters $\widetilde{\alpha}_{nr}$, $\widetilde{\beta}_{nr}$)
initialize $\widetilde{\alpha}_{nr} \leftarrow \widetilde{\alpha}_{nr}^{ini}$, $\widetilde{\beta}_{nr} \leftarrow \widetilde{\beta}_{nr}^{ini}$ for $r = 1, \ldots, R$ (with $r \neq x_n^d$)
initialize $\mathbf{\Pi}_{\mathrm{I}_n}, \boldsymbol{\lambda}_{\mathrm{I}_n}$ from $\mathbf{\Pi}_{\neg n}, \boldsymbol{\lambda}_{\neg n}$
initialize $\widetilde{\mathbf{\Pi}}_{\mathrm{I}_n} \leftarrow \mathbf{\Pi}_{\mathrm{I}_n} + \sum_{r \neq x_n^d} \widetilde{\alpha}_{nr} \mathbf{h}_{nr}^d (\mathbf{h}_{nr}^d)^\top$, $\quad \widetilde{\boldsymbol{\lambda}}_{\mathrm{I}_n} \leftarrow \boldsymbol{\lambda}_{\mathrm{I}_n} + \sum_{r \neq x_n^d} \widetilde{\beta}_{nr} \mathbf{h}_{nr}^d$
repeat
  for $r = 1, \ldots, R$ with $r \neq x_n^d$ (in parallel) **do**
    marginal moments: $v_{nr} \leftarrow (\mathbf{h}_{nr}^d)^\top \widetilde{\mathbf{\Pi}}_{\mathrm{I}_n}^{-1} \mathbf{h}_{nr}^d$, $m_{nr} \leftarrow (\mathbf{h}_{nr}^d)^\top \widetilde{\mathbf{\Pi}}_{\mathrm{I}_n}^{-1} \widetilde{\boldsymbol{\lambda}}_{\mathrm{I}_n}$
    cavity evaluations: $v_{n\neg r} \leftarrow (1/v_{nr} + \widetilde{\alpha}_{nr})^{-1}$, $m_{n\neg r} \leftarrow v_{n\neg r}(m_{nr}/v_{nr} - \widetilde{\beta}_{nr})$
    auxiliary variable:
$$\rho_{nr} \leftarrow \mathcal{N}\left(\frac{m_{n\neg r}}{\sqrt{1+v_{n\neg r}}} | 0, 1\right) / \left(\Phi\left(\frac{m_{n\neg r}}{\sqrt{1+v_{n\neg r}}}\right)\sqrt{1+v_{n\neg r}}\right)$$
    tilted moments:
$$\widehat{v}_{nr} \leftarrow v_{n\neg r} - v_{n\neg r}^2 \left(\rho_{nr}^2 + \rho_{nr} \frac{m_{n\neg r}}{1+v_{n\neg r}}\right),$$
$$\widehat{m}_{nr} \leftarrow m_{n\neg r} + \rho_{nr} v_{n\neg r},$$
$$\widehat{C}_{nr} \leftarrow \Phi\left(\frac{m_{n\neg r}}{\sqrt{1+v_{n\neg r}}}\right)$$
    site updates: $\widetilde{\alpha}_{nr} \leftarrow \eta_{\mathrm{I}}(1/\widehat{v}_{nr} - 1/v_{n\neg r}) + (1-\eta_{\mathrm{I}})\widetilde{\alpha}_{nr}$,
$\widetilde{\beta}_{nr} \leftarrow \eta_{\mathrm{I}}(\widehat{m}_{nr}/\widehat{v}_{nr} - m_{n\neg r}/v_{n\neg r}) + (1-\eta_{\mathrm{I}})\widetilde{\beta}_{nr}$
  end for
  update $\widetilde{\mathbf{\Pi}}_{\mathrm{I}_n} \leftarrow \mathbf{\Pi}_{\mathrm{I}_n} + \sum_{r \neq x_n^d} \widetilde{\alpha}_{nr} \mathbf{h}_{nr}^d (\mathbf{h}_{nr}^d)^\top$, $\widetilde{\boldsymbol{\lambda}}_{\mathrm{I}_n} \leftarrow \boldsymbol{\lambda}_{\mathrm{I}_n} + \sum_{r \neq x_n^d} \widetilde{\beta}_{nr} \mathbf{h}_{nr}^d$
until stopping criterion
for $r = 1, \ldots, R$ with $r \neq x_n^d$ (in parallel) **do**
  compute $\log \widetilde{C}_{nr}$ from (B.9) (see Appendix B)
end for
compute $\log \widehat{Z}_n$ from (B.8) (see Appendix B)
compute $\widehat{\mathbf{\Pi}}_n, \widehat{\boldsymbol{\lambda}}_n$ from $\widetilde{\mathbf{\Pi}}_{\mathrm{I}_n}, \widetilde{\boldsymbol{\lambda}}_{\mathrm{I}_n}$

trying to add new features, the values of the 'old' site parameters can still be used to build the 'new' parameters (extended to account for the new features) $\widetilde{\boldsymbol{\lambda}}_n$ and $\widetilde{\mathbf{\Pi}}_n$.

4.1.3 Inferring the Severity Matrix

So far, we have consider the model without the severity matrix $\mathbf{W}$, being in this case Gibbs sampling suitable to infer the latent feature matrix $\mathbf{Z}$. However, when considering the "full" model in 3.3, in addition to the latent

matrix $\mathbf{Z}$, we need also to infer the severity matrix $\mathbf{W}$, which cannot be done with the Gibbs sampling algorithm. Here, we instead propose an inference algorithm based on Metropolis-Hastings (MH) algorithm, in which we jointly sample z_{nk} and w_{nk} having marginalized the matrices $\mathbf{B}^d$. Since as before, the posterior of $\mathbf{B}^d$ is intractable, we can resort either to the Laplace approximation or to the nested EP algorithm in order to approximately integrate out $\mathbf{B}^d$ to obtain the marginal likelihood $p(\mathbf{X}|\mathbf{W})$. Note that adapting the EP approximation, detailed in Section 4.1.2, to deal with the severity matrix $\mathbf{W}$ can be performed by simply replacing all the references in the nested EP algorithm to the binary matrix $\mathbf{Z}$ by the severity matrix $\mathbf{W}$. The adaptation of the Laplace approximation is also straightforward but some of the equations in Section 4.1.1 change when considering the severity factors.

Our MH based algorithm proceeds iteratively as follows. For each observation $n = 1, \ldots, N$:

- **Step 1**: Jointly sample z_{nk} and w_{nk} for $k = 1, \ldots, K_+$;
- **Step 2**: Consider adding new latent features for the n-th observation, updating K_+ if necessary.

In Step 1, we rely on MH proposing to move from an initial pair (z_{nk}, w_{nk}) to (z^*_{nk}, w^*_{nk}) (jumping from matrices $\mathbf{Z}$ and $\mathbf{W}$ to $\mathbf{Z}^*$ and $\mathbf{W}^*$). Our proposal distribution is

$$q_1(z^*_{nk}, w^*_{nk}|z_{nk}, w_{nk}) = \begin{cases} \delta_1(z^*_{nk})p(w^*_{nk}|z^*_{nk} = 1), & \text{if } z_{nk} = 0, \\ \frac{1}{2}\delta_0(z^*_{nk})\delta_0(w^*_{nk}) + \frac{1}{2}\delta_1(z^*_{nk})p(w^*_{nk}|z^*_{nk} = 1), & \text{if } z_{nk} \neq 0, \end{cases} \tag{4.16}$$

i.e., if $z_{nk} = 0$ we propose to move to $z^*_{nk} = 1$ with w^*_{nk} sampled from

$$p(w^*_{nk}|\gamma_1, \gamma_2, z^*_{nk}) = (1 - z^*_{nk})\delta_0(w^*_{nk}) + z^*_{nk}\text{Beta}(w^*_{nk}|\gamma_1, \gamma_2). \tag{4.17}$$

Otherwise, either a move to $z^*_{nk} = 0$ or to $z^*_{nk} = 1$ (with a value of w^*_{nk} drawn from the Eq. 4.17) is proposed with equal probability. The acceptance probability for the MH step is given by

$$\min\left(1, \frac{p(\mathbf{X}|\mathbf{W}^*)p([\mathbf{Z}^*])p(w^*_{nk}|z^*_{nk})}{p(\mathbf{X}|\mathbf{W})p([\mathbf{Z}])p(w_{nk}|z_{nk})} \frac{q_1(z_{nk}, w_{nk}|z^*_{nk}, w^*_{nk})}{q_1(z^*_{nk}, w^*_{nk}|z_{nk}, w_{nk})}\right), \tag{4.18}$$

where

$$\frac{p([\mathbf{Z}^*])}{p([\mathbf{Z}])} = \begin{cases} 1, & \text{if } z_{nk} = z^*_{nk}, \\ m_{\neg nk}/(N - m_{\neg nk}), & \text{if } z^*_{nk} = 1 \text{ and } z_{nk} = 0, \\ (N - m_{\neg nk})/m_{\neg nk}, & \text{if } z^*_{nk} = 0 \text{ and } z_{nk} = 1, \end{cases} \tag{4.19}$$

being $m_{\neg nk}$ the number of data points (excluding n) which have active the k-th feature, namely, $m_{\neg nk} = \sum_{i \neq n} z_{ik}$. The distribution $p(w_{nk}|z_{nk})$ is given in Eq. 4.17 and, as previously stated, the probabilities $p(\mathbf{X}|\mathbf{W})$ are obtained using either the Laplace approximation or the nested EP algorithm detailed, respectively, in Sections 4.1.1 and 4.1.2.

For Step 2, we need to define κ_n as the number of columns of $\mathbf{Z}$ which are active only in the n-th row, i.e., $\kappa_n = \sum_{k=1}^{\infty} z_{nk} \prod_{i \neq n} (1 - z_{ik})$. Note that, after performing Step 1, the initial value of κ_n is 0 due to the form of Eqs. 4.18 and 4.19. The new value κ_n^* is sampled with a MH step. We include as part of the proposal the corresponding new values of the severity matrix, i.e., a $1 \times \kappa_n^*$ vector denoted by $\boldsymbol{\omega}_n^*$. Therefore, we propose to jump from a initial value of κ_n and $\boldsymbol{\omega}_n$ to κ_n^* and $\boldsymbol{\omega}_n^*$, where the latter variables are drawn from the proposal distribution

$$q_2(\kappa_n^*, \boldsymbol{\omega}_n^*) = q_2(\kappa_n^*) q_2(\boldsymbol{\omega}_n^* | \kappa_n^*). \tag{4.20}$$

We make $q_2(\boldsymbol{\omega}_n^*|\kappa_n^*)$ equal to the prior, i.e., $q_2(\boldsymbol{\omega}_n^*|\kappa_n^*) = \prod_{k'=1}^{\kappa_n^*} p(\omega_{nk'}^*|z_{nk'}^* = 1)$, and $q_2(\kappa_n^*)$ is chosen as in [41], namely,

$$q_2(\kappa_n^*) = (1 - \pi)\text{Poisson}\left(\kappa_n^* \,|\alpha\lambda/N\right) + \pi\delta_1(\kappa_n^*), \tag{4.21}$$

where we set $\lambda = N/2$ and $\pi = 0.2$. The move is accepted with probability

$$\min\left(1, \frac{p(\mathbf{X}|\mathbf{W}^*)}{p(\mathbf{X}|\mathbf{W})} \frac{(\alpha/N)^{\kappa_n^*}}{\kappa_n^*!} \frac{q_2(\kappa_n)}{q_2(\kappa_n^*)}\right). \tag{4.22}$$

4.1.4 Laplace Approximation vs. Expectation Propagation

In order to evaluate the performance of our model and inference algorithms, we generate two synthetic datasets and perform comparisons between a latent feature model with:

(i) On/off hidden variables and inference based on Gibbs sampling combined with the Laplace approximation described in Section 4.1.1 (denoted by "On/Off+Lap.");

(ii) On/off hidden variables and inference based on Gibbs sampling and the nested EP approximation described in Section 4.1.2 ("On/Off+EP");

(iii) Continuous hidden variables in $[0, 1]$ and inference based on MH steps and the nested EP approximation, i.e., the algorithm in Section 4.1.3 ("Sev.+EP").

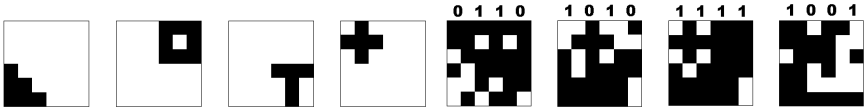


Figure 4.1: Toy example 1. (a) Base images. (b) Four observation examples. The numbers above each figure indicate which features are present in that image.

We generate binary-valued observation matrices $\mathbf{X}$, with $N = 100$ black-and-white images with dimensionality $D = 36$, that are built differently for each of the two datasets.

Toy Example 1: In this example, each observation $\mathbf{x}_n$ is a combination of four latent black-and-white base images that can be present or absent with probability 0.5 independently of each other, i.e., $z_{nk} = 1$ with probability 0.5. Each white pixel in the composite image becomes black with probability 0.5, while black pixels remain black. We plot in Figure 4.1 the four base images and four observation examples.

Toy Example 2: This example is similar to the previous one but we introduce a latent auxiliary matrix $\mathbf{A}$ to generate observations. As before, we assume four latent features that become active with probability 0.5, but we also generate a $N \times 4$ matrix $\mathbf{A}$, whose elements a_{nk} are Beta$(2, 1)$ distributed. In this set-up, we divide each image into four disjoint regions of 9 pixels, each modelled by one of the latent features. Each of the 9 pixels in the observation n corresponding to feature k are set to black with probability $0.5 + 0.5a_{nk}$ if $z_{nk} = 1$, or with probability 0.5 otherwise.

Validation: In order to compare the three methods, we average over 5 independent realizations of the two synthetic datasets the following scores:

- Approximate marginal log-likehood (Log-lik).
- Kullback-Leibler divergence (D_{KL}) between the true and the inferred probability of the observation matrix. We compute the inferred probability using the mean of the approximate posterior of $\mathbf{B}^d$ and the sample of the latent feature matrix $\mathbf{Z}$ (or $\mathbf{W}$, if available).

In Tables 4.1 and 4.2 we show the results for the two synthetic datasets. Note that the obtained values of the average log-likelihood are similar for the three considered methods (no significant statistical differences are found) in both examples. However, we can observe significant differences in terms of the Kullback-Leibler divergence, for which the model with severity factors

combined with the EP inference provides the best results in both examples. Additionally, in Toy Example 1, since the generative model considers binary latent variables (instead of continuous), both the "On/Off+EP" and the "Sev.+EP" methods provide similar results. Hence, in agreement with previous works [45, 64], we observe that the EP algorithm provides better estimates of the marginal likelihood than the Laplace approximation, and the severity factors included in the Full model in Figure 3.3 lead to a more flexible and expressive model (and the corresponding inference algorithm) that is able to better explain diverse databases.

	On/Off+Lap.	On/Off+EP	Sev.+EP
Log-lik	$-1,943$	$-2,001$	$-1,948$
D_{KL}	497.15	354.92	347.11

Table 4.1: Results for the Toy Example 1.

	On/Off+Lap.	On/Off+EP	Sev.+EP
Log-lik	$-2,122$	$-2,233$	$-2,151$
D_{KL}	524.16	372.10	353.15

Table 4.2: Results for the Toy Example 2.

4.2 Variational Inference

Variational inference provides a complementary (and less expensive in terms of computational complexity) alternative to MCMC methods as a general source of approximation methods for inference in large-scale statistical models [36]. The main idea behind variational inference is to approximate the distribution by a tractable distribution and find the optimal parameters by minimizing the KL divergence between both distributions. In this chapter, we adapt the infinite variational approach for the standard linear-Gaussian IBP model, introduced in [19], to the multinomial-logit IBP model introduced in Chapter 3. This approach assumes the (truncated) stick-breaking construction for the IBP detailed in Section 2.2.1, which bounds the number of columns of the IBP matrix by a finite (but large enough) value, K. Then, in the truncated stick-breaking process, $\omega_k = \prod_{i=1}^{k} v_i$ for $k \leq K$, and zero otherwise.

The hyperparameters of the model are grouped in the set $\mathcal{H} = \{\alpha, \sigma_B^2\}$ and, similarly, $\Psi = \{\mathbf{Z}, \mathbf{B}^1, \ldots, \mathbf{B}^D, \mathbf{b}_0^1, \ldots, \mathbf{b}_0^D, v_1, \ldots, v_K\}$ denotes the set

of unobserved variables in the model where, for clarity, we explicitly account for the bias terms $\mathbf{b}_0^d$. Under the truncated stick-breaking construction for the IBP, the joint probability distribution over all the variables $p(\Psi, \mathbf{X}|\mathcal{H})$ can be factorized as

$$p(\Psi, \mathbf{X}|\mathcal{H}) = \prod_{k=1}^{K}\left(p(v_k|\alpha)\prod_{n=1}^{N}p(z_{nk}|\{v_i\}_{i=1}^{k})\right)\prod_{d=1}^{D}\left(p(\mathbf{b}_0^d|\sigma_B^2)\prod_{k=1}^{K}p(\mathbf{b}_{k\cdot}^d|\sigma_B^2)\right) \times \prod_{n=1}^{N}\prod_{d=1}^{D}p(x_n^d|\mathbf{z}_n, \mathbf{B}^d, \mathbf{b}_0^d),$$

where $\mathbf{b}_{k\cdot}^d$ is the k^{th} row of matrix $\mathbf{B}^d$.

We approximate $p(\Psi|\mathbf{X}, \mathcal{H})$ with the variational distribution $q(\Psi)$ given by

$$q(\Psi) = \prod_{k=1}^{K}\left(q(v_k|\tau_{k1}, \tau_{k2})\prod_{n=1}^{N}q(z_{nk}|\nu_{nk})\right)\prod_{k=0}^{K}\prod_{r=1}^{R}\prod_{d=1}^{D}q(b_{kr}^d|\phi_{kr}^d, (\sigma_{kr}^d)^2),$$

where the terms b_{kr}^d stand for the elements of matrix $\mathbf{B}^d$, and

$$\begin{aligned} &q(v_k|\tau_{k1}, \tau_{k2}) = \text{Beta}(\tau_{k1}, \tau_{k2}),\\ &q(b_{kr}^d|\phi_{kr}^d, (\sigma_{kr}^d)^2) = \mathcal{N}(\phi_{kr}^d, (\sigma_{kr}^d)^2),\\ &q(z_{nk}|\nu_{nk}) = \text{Bernoulli}(\nu_{nk}). \end{aligned}$$

Inference involves optimizing the variational parameters of $q(\Psi)$ to minimize the Kullback-Leibler divergence from $q(\Psi)$ to $p(\Psi|\mathbf{X}, \mathcal{H})$, i.e., $D_{KL}(q||p)$. This optimization is equivalent to maximizing a lower bound on the evidence $p(\mathbf{X}|\mathcal{H})$, which can be computed as

$$\begin{aligned} \log p(\mathbf{X}|\mathcal{H}) &= \mathbb{E}_q\left[\log p(\Psi, \mathbf{X}|\mathcal{H})\right] + H[q] + D_{KL}(q||p)\\ &\geqslant \mathbb{E}_q\left[\log p(\Psi, \mathbf{X}|\mathcal{H})\right] + H[q], \end{aligned} \tag{4.23}$$

where $\mathbb{E}_q[\cdot]$ denotes the expectation with respect to the distribution $q(\Psi)$, $H[q]$ is the entropy of distribution $q(\Psi)$, and

$$\begin{aligned} &\mathbb{E}_q\left[\log p(\Psi, \mathbf{X}|\mathcal{H})\right]\\ &= \sum_{k=1}^{K}\mathbb{E}_q\left[\log p(v_k|\alpha)\right] + \sum_{d=1}^{D}\sum_{k=1}^{K}\mathbb{E}_q\left[\log p(\mathbf{b}_{k\cdot}^d|\sigma_B^2)\right] + \sum_{d=1}^{D}\mathbb{E}_q\left[\log p(\mathbf{b}_0^d|\sigma_B^2)\right]\\ &\quad + \sum_{k=1}^{K}\sum_{n=1}^{N}\mathbb{E}_q\left[\log p(z_{nk}|\{v_i\}_{i=1}^{k})\right] + \sum_{n=1}^{N}\sum_{d=1}^{D}\mathbb{E}_q\left[\log p(x_n^d|\mathbf{z}_n, \mathbf{B}^d, \mathbf{b}_0^d)\right]. \end{aligned} \tag{4.24}$$

The derivation of the lower bound in (4.23) is straightforward, with the exception of the terms $\mathbb{E}_q\left[\log p(z_{nk}|\{v_i\}_{i=1}^k)\right]$ and $\mathbb{E}_q\left[\log p(x_n^d|\mathbf{z}_n, \mathbf{B}^d, \mathbf{b}_0^d)\right]$ in (4.24), which have no closed-form solution, so we instead bound them. Deriving these bounds leads to a new bound $\mathcal{L}(\mathcal{H}, \mathcal{H}_q)$ (that can be written in closed-form) such that $\log p(\mathbf{X}|\mathcal{H}) \geq \mathcal{L}(\mathcal{H}, \mathcal{H}_q)$, being $\mathcal{H}_q$ the full set of variational parameters. The final expression for $\mathcal{L}(\mathcal{H}, \mathcal{H}_q)$, as well as the details on the derivation of the bound, are provided in Appendix C.1.

In order to maximize the lower bound $\mathcal{L}(\mathcal{H}, \mathcal{H}_q)$, we need to optimize with respect to the value of the variational parameters. To this end, we can iteratively maximize the bound with respect to each variational parameter by taking the derivative of $\mathcal{L}(\mathcal{H}, \mathcal{H}_q)$ and setting it to zero. This procedure readily leads to the following fixed-point equations:

1. For the variational Beta distribution $q(v_k|\tau_{k1}, \tau_{k2})$,

$$\tau_{k1} = \alpha + \sum_{m=k}^{K}\left(\sum_{n=1}^{N}\nu_{nm}\right) + \sum_{m=k+1}^{K}\left(N - \sum_{n=1}^{N}\nu_{nm}\right)\left(\sum_{i=k+1}^{m}\lambda_{mi}\right),$$

$$\tau_{k2} = 1 + \sum_{m=k}^{K}\left(N - \sum_{n=1}^{N}\nu_{nm}\right)\lambda_{mk}.$$

2. For the Bernoulli distribution $q(z_{nk}|\nu_{nk})$,

$$\nu_{nk} = \frac{1}{1 + \exp(-A_{nk})},$$

where

$$\begin{aligned}
&A_{nk} \\
&= \sum_{i=1}^{k}\left[\psi(\tau_{i1}) - \psi(\tau_{i1} + \tau_{i2})\right] - \left[\sum_{m=1}^{k}\lambda_{km}\psi(\tau_{m2}) + \sum_{m=1}^{k-1}\left(\sum_{n=m+1}^{k}\lambda_{kn}\right)\psi(\tau_{m1})\right. \\
&\quad \left. - \sum_{m=1}^{k}\left(\sum_{n=m}^{k}\lambda_{kn}\right)\psi(\tau_{m1} + \tau_{m2}) - \sum_{m=1}^{k}\lambda_{km}\log(\lambda_{km})\right] \\
&+ \sum_{d=1}^{D}\left(\phi^d_{kx_n^d} - \xi_{nd}\sum_{r=1}^{R}\left[\exp\left(\phi^d_{0r} + \frac{1}{2}(\sigma^d_{0r})^2\right)\left(1 - \exp\left(\phi^d_{kr} + \frac{1}{2}(\sigma^d_{kr})^2\right)\right)\times\right.\right. \\
&\qquad \left.\left.\times\prod_{k'\neq k}\left(1 - \nu_{nk'} + \nu_{nk'}\exp\left(\phi^d_{k'r} + \frac{1}{2}(\sigma^d_{k'r})^2\right)\right)\right]\right),
\end{aligned}$$

and $\psi(\cdot)$ stands for the digamma function [3, p. 258–259].

3. For the feature assignments, which are Bernoulli distributed given the feature probabilities, we have lower bounded $\mathbb{E}_q\left[\log p(z_{nk}|\{v_i\}_{i=1}^k)\right]$ by using the multinomial approach in [19] (see Appendix C.1 for further details). This approximation introduces the auxiliary multinomial distribution $\boldsymbol{\lambda}_k = [\lambda_{k1}, \ldots, \lambda_{kk}]$, where each λ_{ki} can be updated as

$$\lambda_{ki} \propto \exp\left(\psi(\tau_{i2}) + \sum_{m=1}^{i-1}\psi(\tau_{m1}) - \sum_{m=1}^{i}\psi(\tau_{m1}+\tau_{m2})\right),$$

where the proportionality ensures that $\boldsymbol{\lambda}_k$ is a valid distribution.

4. The maximization with respect to the variational parameters ϕ_{kr}^d, ϕ_{0r}^d, $(\sigma_{kr}^d)^2$, and $(\sigma_{0r}^d)^2$ has no analytical solution, and therefore, we need to resort to a numerical method to find the maximum, such as Newton's method or conjugate gradient algorithm, for which the first and the second derivatives[1] (given in Appendix C.2) are required.

5. Finally, we lower bound the likelihood term $\mathbb{E}_q\left[\log p(x_n^d|\mathbf{z}_n, \mathbf{B}^d, \mathbf{b}_0^d)\right]$ by resorting to a first-order Taylor series expansion around the auxiliary variables ξ_{nd}^{-1} for $n = 1, \ldots, N$ and $d = 1, \ldots, D$ (see Appendix C.1 for further details), which are optimized by the expression

$$\xi_{nd} = \left[\sum_{r=1}^{R}\exp\left(\phi_{0r}^d + \frac{1}{2}(\sigma_{0r}^d)^2\right)\prod_{k=1}^{K}\left(1 - \nu_{nk} + \nu_{nk}\exp\left(\phi_{kr}^d + \frac{1}{2}(\sigma_{kr}^d)^2\right)\right)\right]^{-1}.$$

[1] Note that the second derivatives are strictly negative and, therefore, the maximum with respect to each parameter is unique.

Chapter 5

IBP for Heterogeneous Databases

In previous chapters, we have focused on categorical data because most of the available data in psychiatry and, more specifically, in the NESARC database are categorical, i.e., they take values in a finite unordered set, e.g., {'yes', 'no', 'unknown'}. However, the database contains other non-categorical attributes, frequently related to the subject background, such us age, highest grade level completed in school, income, etc., that can also be relevant for the manifestation of psychiatric disorders. For instance, the number of consumed alcoholic drinks per day, which can be modeled as count data, appears as a key variable to detect those subjects that suffer from alcohol use disorder.

In this chapter, we aim at providing a general model that allows handling all the available information about the subjects. In particular, we propose a general observation model for the IBP that accounts for heterogeneous data, where the attributes describing each subject can be either discrete (categorical, ordinal and count), continuous (real-valued and positive real-valued) or mixed variables. The proposed model keeps the properties of conjugate models and allows us to derive an efficient inference algorithm that scales linearly with the number of observations. In the literature, we find that latent feature model approaches usually assume homogeneous databases with either real [66, 67, 84] or categorical data [46], and only a few works consider heterogeneous data, such as mixed real and categorical data [68]. However, up to our knowledge, there are no general latent feature models to directly deal with heterogeneous databases.

5.1 Model Description

Let us assume a database with N objects, where each object is defined by D attributes. We can store the data in an $N \times D$ observation matrix $\mathbf{X}$, in which each D-dimensional row vector is denoted by $\mathbf{x}_n = [x_n^1, \ldots, x_n^D]$ and each entry is denoted by x_n^d. We consider that the column vectors $\mathbf{x}^d$ (i.e., each dimension in the observation matrix $\mathbf{X}$) may contain the following types of data:

- Continuous variables:
  1. Real-valued, i.e., $x_n^d \in \Re$.
  2. Positive real-valued, i.e., $x_n^d \in \Re_+$.
- Discrete variables:
  1. Categorical data, i.e., x_n^d takes values in a finite unordered set, e.g., $x_n^d \in \{$'blue', 'red', 'black'$\}$.
  2. Ordinal data, i.e., x_n^d takes values in a finite ordered set, e.g., $x_n^d \in \{$'never', 'sometimes', 'often', 'usually', 'always'$\}$.
  3. Count data, i.e., $x_n^d \in \{0, 1, 2, \ldots, \infty\}$.

As proposed in Chapter 3, we assume that each observation x_n^d can be explained by a K-length vector of latent variables associated to the n-th data point $\mathbf{z}_n = [z_{n1}, \ldots, z_{nK}]$ and a weighting vector $\mathbf{B}^d = [b_1^d, \ldots, b_K^d]$ (being K the number of latent variables), whose elements b_k^d weight the contribution of k-th the latent feature to the d-th dimension of the observation matrix $\mathbf{X}$. The binary feature vectors $\mathbf{z}_n$ are stored in the $N \times K$ matrix $\mathbf{Z}$, which follows an IBP with concentration parameter α (i.e., $\mathbf{Z} \sim \text{IBP}(\alpha)$), and the weighting vectors $\mathbf{B}^d$ are Gaussian distributed with zero mean and covariance matrix $\sigma_B^2 \mathbf{I}_K$. For convenience, $\mathbf{z}_n$ is a K-length row vector, while $\mathbf{B}^d$ a K-length column vector.

To accommodate for all kinds of observed random variables described above, we introduce an auxiliary Gaussian variable y_n^d, such that when conditioned on the auxiliary variables, the latent variable model behaves as a standard IBP with Gaussian observations [27]. In particular, we place over y_n^d a Gaussian distribution with mean $\mathbf{z}_n \mathbf{B}^d$ and variance σ_y^2, i.e.,

$$p(y_n^d | \mathbf{z}_n, \mathbf{B}^d) = \mathcal{N}(y_n^d | \mathbf{z}_n \mathbf{B}^d, \sigma_y^2),$$

and assume that there exists a transformation function over the variables y_n^d to obtain the observations x_n^d, mapping the real line $\Re$ into the observation space. The resulting generative model is shown in Figure 5.1, where $\mathbf{Z}$ is

the IBP latent matrix, and $\mathbf{Y}^d$ and $\mathbf{B}^d$ contain, respectively, the auxiliary Gaussian variables y_n^d and the weighting factors b_k^d for the d-dimension of the data. Additionally, Ψ^d denotes the set of auxiliary random variables needed to obtain the observation vector $\mathbf{x}^d$ given $\mathbf{Y}^d$, and $\mathcal{H}^d$ contains the hyper-parameters associated to the random variables in Ψ^d.

This model assumes that the observations x_n^d are independent given the latent matrix $\mathbf{Z}$, the weighting factors $\mathbf{B}^d$ and the auxiliary variables Ψ^d. Therefore, the likelihood can be factorized as

$$p(\mathbf{X}|\mathbf{Z}, \{\mathbf{B}^d, \Psi^d\}_{d=1}^D) = \prod_{d=1}^{D} p(\mathbf{x}^d|\mathbf{Z}, \mathbf{B}^d, \Psi^d) = \prod_{d=1}^{D}\prod_{n=1}^{N} p(x_n^d|\mathbf{z}_n, \mathbf{B}^d, \Psi^d).$$

Note that, if we assume Gaussian observations and set $\mathbf{Y}^d = \mathbf{x}^d$, this model resembles the standard IBP with Gaussian observations [27]. In addition, conditioned on the variables $\mathbf{Y}^d$, we can infer the latent matrix $\mathbf{Z}$ as in the standard IBP. We also remark that auxiliary Gaussian variables to link the latent model with the observations have been previously used in Gaussian processes for multi-class classification [25] and for ordinal regression [15]. However, up to our knowledge, this simple approach has not been used to account for mixed continuous and discrete data, existing a lack of work in this field.

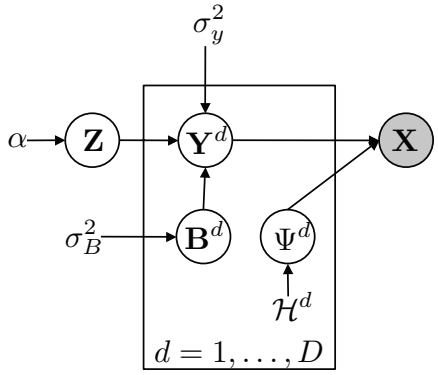


Figure 5.1: Generalized IBP for mixed continuous and discrete observations.

5.1.1 Likelihood Functions

Now, we define the set of transformations that map from the Gaussian variables y_n^d to the corresponding observations x_n^d. We assume that each

column in matrix $\mathbf{X}$ may contain any of the discrete or continuous variables detailed above, provide a likelihood function for each kind of data and, in turn, also a likelihood function for mixed data.

Real-valued Data. In this case, we assume that $\mathbf{x}^d = \mathbf{Y}^d$ in the model in Figure 5.1 and consider the standard approach when dealing with real-valued observations, which consist of assuming a Gaussian likelihood function. In particular, as in the standard linear-Gaussian IBP [27], we assume that each observation x_n^d is distributed as

$$p(x_n^d|\mathbf{z}_n, \mathbf{B}^d) = \mathcal{N}(x_n^d|\mathbf{z}_n\mathbf{B}^d, \sigma_y^2).$$

Positive Real-valued Data. In order to obtain positive real-valued observations, i.e., $x_n^d \in \Re_+$, we apply a transformation over y_n^d that maps from the real numbers to the positive real numbers, i.e.,

$$x_n^d = f(y_n^d + u_n^d),$$

where u_n^d is a Gaussian noise variable with variance σ_u^2, and $f : \Re \to \Re_+$ is a differentiable function. By change of variables, we obtain the likelihood function for positive real-valued observations as

$$\begin{aligned} &p(x_n^d|\mathbf{z}_n, \mathbf{B}^d) \\ &= \frac{1}{\sqrt{2\pi(\sigma_y^2 + \sigma_u^2)}} \exp\left\{-\frac{1}{2(\sigma_y^2 + \sigma_u^2)}(f^{-1}(x_n^d) - \mathbf{z}_n\mathbf{B}^d)^2\right\} \left|\frac{d}{dx_n^d} f^{-1}(x_n^d)\right|, \end{aligned} \tag{5.1}$$

where $f^{-1} : \Re_+ \to \Re$ is the inverse function of the transformation $f(\cdot)$, i.e, $f^{-1}(f(v)) = v$. Note that, in this case, we make use of the Gaussian variable u_n^d to obtain x_n^d from y_n^d, and therefore, $\Psi^d = \{u_d^n\}_{n=1}^N$ and $\mathcal{H}^d = \sigma_u^2$.

Categorical Data. Now, we account for categorical observations, i.e., each observation x_n^d can take values in the unordered index set $\{1, \ldots, R_d\}$. Hence, assuming a multinomial-probit model, we can write

$$x_n^d = \arg \max_{r \in \{1,\ldots,R_d\}} y_{nr}^d, \tag{5.2}$$

being $y_{nr}^d \sim \mathcal{N}(y_{nr}^d|\mathbf{z}_n\mathbf{b}_r^d, \sigma_y^2)$ where $\mathbf{b}_r^d$ denotes the K-length weighting vector, in which each entry b_{kr}^d weights the influence of the k-th feature for the observation x_n^d taking value r. Note that, under this likelihood model, since we have a Gaussian auxiliary variable y_{nr}^d and a weighting factor b_{kr}^d

for each possible value of the observation $r \in \{1, \ldots, R_d\}$, we need to gather all the weighting factors b_{kr}^d in a $K \times R_d$ matrix $\mathbf{B}^d$, and all the Gaussian auxiliary variables y_{nr}^d in the $N \times R_d$ matrix $\mathbf{Y}^d$.

Under this observation model, we can write $y_{nr}^d = \mathbf{z}_n \mathbf{b}_r^d + u_{nr}^d$, where u_{nr}^d is a Gaussian noise variable with variance σ_y^2, and therefore, we can obtain the probability of each element x_n^d taking value $r \in \{1, \ldots, R_d\}$ as [25]

$$p(x_n^d = r | \mathbf{z}_n, \mathbf{B}^d) = \mathbb{E}_{p(u)} \left[\prod_{\substack{j=1 \\ j \neq r}}^{R_d} \Phi\Big(u + \mathbf{z}_n (\mathbf{b}_r^d - \mathbf{b}_j^d)\Big) \right], \tag{5.3}$$

where subscript r in $\mathbf{b}_r^d$ indicates the column in $\mathbf{B}^d$ ($r \in \{1, \ldots, R_d\}$), $\Phi(\cdot)$ denotes the cumulative density function of the standard normal distribution and $\mathbb{E}_{p(u)}[\cdot]$ denotes expectation with respect to the distribution $p(u) = \mathcal{N}(0, \sigma_y^2)$. Note that this likelihood model coincides with the multinomial-probit likelihood introduced in Chapter 3, but here we exploit the underlying structure of the probit model to obtain Eq. 5.2 which, opposite to Eq. 5.3, allows us to derive an exact collapsed Gibbs sampler by conditioning on the auxiliary variables y_{nr}^d.

Ordinal Data. Consider ordinal data, in which each element x_n^d takes values in the ordered index set $\{1, \ldots, R_d\}$. Then, assuming an ordered probit model, we can write

$$x_n^d = \begin{cases} 1 & \text{if } y_n^d \leq \theta_1^d \\ 2 & \text{if } \theta_1^d < y_n^d \leq \theta_2^d \\ & \vdots \\ R_d & \text{if } \theta_{R_d-1}^d < y_n^d \end{cases} \tag{5.4}$$

where again y_n^d is Gaussian distributed with mean $\mathbf{z}_n \mathbf{B}^d$ and variance σ_y^2, and θ_r^d for $r \in \{1, \ldots, R_d - 1\}$ are the thresholds that divide the real line into R_d regions. We assume the thresholds θ_r^d are sequentially generated from the truncated Gaussian distribution $\theta_r^d \propto \mathcal{N}(\theta_r^d | 0, \sigma_\theta^2) \mathbb{I}(\theta_r^d > \theta_{r-1}^d)$, where $\theta_0^d = -\infty$ and $\theta_{R_d}^d = +\infty$. As opposed to the categorical case, now we have a unique weighting vector $\mathbf{B}^d$ and a unique Gaussian variable y_n^d for each observation x_n^d. Hence, the value of x_n^d is determined by the region in which y_n^d falls.

Under the ordered probit model [15], the probability of each element x_n^d taking value $r \in \{1, \ldots, R_d\}$ can be written as

$$p(x_n^d = r | \mathbf{z}_n, \mathbf{B}^d) = \Phi\left(\frac{\theta_r^d - \mathbf{z}_n \mathbf{B}^d}{\sigma_y} \right) - \Phi\left(\frac{\theta_{r-1}^d - \mathbf{z}_n \mathbf{B}^d}{\sigma_y} \right). \tag{5.5}$$

Let us remark that, if the d-dimension of the observation matrix contains ordinal data, the set of auxiliary variables reduces to the thresholds $\Psi^d = \{\theta_1^d, \ldots, \theta_{R_d-1}^d\}$ and $\mathcal{H}^d = \sigma_\theta^2$.

Count Data. In count data, each observation x_n^d takes non-negative integer values, i.e., $x_n^d \in \{0, 1, 2, \ldots, \infty\}$. Then, we assume

$$x_n^d = \lfloor f(y_n^d) \rfloor, \tag{5.6}$$

where $\lfloor v \rfloor$ returns the floor of v, that is the largest integer that does not exceed v, and $f : \Re \to \Re_+$ is a differentiable function that maps from the real numbers to the positive real numbers. We can therefore write the likelihood function as

$$p(x_n^d|\mathbf{z}_n, \mathbf{B}^d) = \Phi\left(\frac{f^{-1}(x_n^d + 1) - \mathbf{z}_n\mathbf{B}^d}{\sigma_y}\right) - \Phi\left(\frac{f^{-1}(x_n^d) - \mathbf{z}_n\mathbf{B}^d}{\sigma_y}\right) \tag{5.7}$$

where $f^{-1} : \Re_+ \to \Re$ is the inverse function of the transformation $f(\cdot)$.

5.2 Inference Algorithm

In this section, we describe our algorithm for inferring the latent variables given the observation matrix. Under our model, detailed in Section 5.1, the probability distribution over the observation matrix is fully characterized by the latent matrices $\mathbf{Z}$ and $\{\mathbf{B}^d\}_{d=1}^D$ (as well as the auxiliary variables Ψ^d).

We use Markov Chain Monte Carlo (MCMC) methods, which have been broadly applied to infer the IBP matrix (see, e.g., in [27, 95, 83]). The proposed inference algorithm is summarized in Algorithm 3. This algorithm exploits the information in the available data to learn the similarities among the objects (captured in our model by the latent feature matrix $\mathbf{Z}$), and how these latent features show up in the attributes that describe the objects (captured in our model by $\mathbf{B}^d$).

In Algorithm 3, we first need to update the latent IBP matrix $\mathbf{Z}$. Note that conditioned on $\{\mathbf{Y}^d\}_{d=1}^D$, both the latent matrix $\mathbf{Z}$ and the weighting matrices $\{\mathbf{B}^d\}_{d=1}^D$ are independent of the observation matrix $\mathbf{X}$. Additionally, since $\{\mathbf{B}^d\}_{d=1}^D$ and $\{\mathbf{Y}^d\}_{d=1}^D$ are Gaussian distributed, we can analytically marginalize out the weighting matrices $\{\mathbf{B}^d\}_{d=1}^D$ to obtain $p(\{\mathbf{Y}^d\}_{d=1}^D|\mathbf{Z})$. Therefore, to infer the IBP matrix $\mathbf{Z}$, we can apply the collapsed Gibbs sampler which presents better mixing properties than the uncollapsed Gibbs sampler and, in consequence, is the standard method of choice in the context of the standard linear-Gaussian IBP [27]. However,

Algorithm 3 Inference Algorithm.

Require: $\mathbf{X}$
Ensure: initialize $\mathbf{Z}$ and $\{\mathbf{Y}^d\}_{d=1}^D$
1: **for** each iteration **do**
2: Update $\mathbf{Z}$ given $\{\mathbf{Y}^d\}_{d=1}^D$.
3: **for** $d = 1, \ldots, D$ **do**
4: Sample $\mathbf{B}^d$ given $\mathbf{Z}$ and $\mathbf{Y}^d$ according to (5.8).
5: Sample $\mathbf{Y}^d$ given $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{B}^d$ as shown in Section 5.2.2.
6: Sample Ψ^d if needed as shown in Section 5.2.2.
7: **end for**
8: **end for**
Output: $\mathbf{Z}$, $\{\mathbf{B}^d\}_{d=1}^D$ and $\{\Psi^d\}_{d=1}^D$

this algorithm suffers from a high computational cost (being the complexity per iteration cubic with the number of data points N), which is prohibitive when dealing with large databases. In order to solve this limitation, we instead resort to the accelerated Gibbs sampler [18]. This algorithm allows us to integrate out the weighting factors in $\{\mathbf{B}^d\}_{d=1}^D$ while keeping linear complexity with the number of datapoints. The accelerated Gibbs sampler is detailed in Section 5.2.1.

Second, we need to sample the weighting factors in $\mathbf{B}^d$, which is a $K \times R_d$ matrix in the case of categorical attributes, and a K-length column vector, otherwise. We denote the r-th column vector of $\mathbf{B}^d$ by $\mathbf{b}_r^d$, where $r \in \{1, \ldots, R_d\}$ when dealing with categorical attributes, and $r = 1$ otherwise. The posterior distributions over the weighting vectors are given by

$$p(\mathbf{b}_r^d | \mathbf{y}_r^d, \mathbf{Z}) = \mathcal{N}(\mathbf{b}_r^d | \mathbf{P}^{-1} \boldsymbol{\lambda}_r^d, \mathbf{P}^{-1}), \tag{5.8}$$

where $\mathbf{P} = \mathbf{Z}^\top \mathbf{Z} + 1/\sigma_B^2 \mathbf{I}_k$ and $\boldsymbol{\lambda}_r^d = \mathbf{Z}^\top \mathbf{y}_r^d$. Note that the covariance matrix $\mathbf{P}^{-1}$ depend neither on the dimension d nor on r, so we only need to invert the $K \times K$ matrix $\mathbf{P}$ once at each iteration. We describe in Section 5.2.1 how to efficiently compute $\mathbf{P}$ after changes in the $\mathbf{Z}$ matrix by rank one updates, without the need of computing the matrix product $\mathbf{Z}^\top \mathbf{Z}$.

Once we have updated $\mathbf{Z}$ and $\mathbf{B}^d$, we sample each element in $\mathbf{Y}^d$ from the posterior distribution $p(y_{nr}^d | x_n^d, \mathbf{z}_n, \mathbf{B}^d)$. The expression of the posterior distribution $p(y_{nr}^d | x_n^d, \mathbf{z}_n, \mathbf{B}^d)$ under each likelihood model in Section 5.1.1 is provided in Section 5.2.2. Finally, we sample the auxiliary variables in Ψ^d from their posterior distribution (detailed in Section 5.2.2) if necessary. This two latter steps involve, in the worst case, sampling from a doubly truncated univariate normal distribution (see the Section 5.2.2 for further

details), for which we make use of the algorithm in [65].

5.2.1 Accelerated Gibbs Sampler

In [18], the authors presented a linear-time accelerated Gibbs sampler for conjugate IBP models that effectively marginalized over the latent factors. The per-iteration complexity of this algorithm is $\mathcal{O}(N(K^2 + KD))$, which is comparable to the uncollapsed linear-Gaussian IBP sampler that has per-iteration complexity $\mathcal{O}(NDK^2)$ but does not marginalize over the weighting factors, and as a result, presents slower convergence rate.

This algorithm exploits the Bayes rule to avoid the cubic complexity with N due to the computation of the marginal likelihood in the Collapsed Gibbs sampler. In particular, it applies the Bayes rule to obtain the probability of each element in the latent feature matrix $\mathbf{Z}$ being active as

$$p(z_{nk} = 1|\{\mathbf{Y}^d\}_{d=1}^D, \mathbf{Z}_{\neg nk}) \propto \frac{m_{\neg n,k}}{N} \prod_{d=1}^{D} \prod_{r=1}^{S_d} \int_{\mathbf{b}_r^d} p(y_{nr}^d|\mathbf{z}_n, \mathbf{b}_r^d) p(\mathbf{b}_r^d|\mathbf{y}_{\neg nr}^d \mathbf{Z}_{\neg n}) d\mathbf{b}_r^d, \tag{5.9}$$

where S_d is the number of columns in matrices $\mathbf{Y}^d$ and $\mathbf{B}^d$ (being S_d the number of categories R_d for those dimension d that contains categorical attributes, and $S_d = 1$ otherwise), $\mathbf{Z}_{\neg n}$ corresponds to matrix $\mathbf{Z}$ after removing the the n-th row, the vector $\mathbf{y}_{\neg nr}^d$ is the $r-$th column of matrix $\mathbf{Y}^d$ without the element y_{nr}^d, and $p(\mathbf{b}_r^d|\mathbf{x}_{\neg n}^d, \mathbf{Z}_{\neg n})$ is the posterior of $\mathbf{b}_r^d$ computed without taking the n-th datapoint into account, i.e.,

$$p(\mathbf{b}_r^d|\mathbf{y}_{\neg nr}^d, \mathbf{Z}_{\neg n}) = \mathcal{N}(\mathbf{b}_r^d|\mathbf{P}_{\neg n}^{-1}\boldsymbol{\lambda}_{\neg nr}^d, \mathbf{P}_{\neg n}^{-1}), \tag{5.10}$$

where $\mathbf{P}_{\neg n} = \mathbf{Z}_{\neg n}^\top \mathbf{Z}_{\neg n} + 1/\sigma_B^2 \mathbf{I}_K$ and $\boldsymbol{\lambda}_{\neg ny}^d = \mathbf{Z}_{\neg n}^\top \mathbf{y}_{\neg nr}^d$ are the natural parameters of the Gaussian distribution.

Note that, opposite to the notation in [18], we here resort to the natural parameters for the Gaussian distribution over the posterior of $\mathbf{b}_r^d$ instead of the mean and the covariance matrix. This formulation allows us to compute the full posterior over the weighting factors as

$$p(\mathbf{b}_r^d|\mathbf{y}_r^d, \mathbf{Z}) = \mathcal{N}(\mathbf{b}_r^d|\mathbf{P}^{-1}\boldsymbol{\lambda}_r^d, \mathbf{P}^{-1}), \tag{5.11}$$

where $\mathbf{P} = \mathbf{P}_{\neg n} + \mathbf{z}_n^\top \mathbf{z}_n$ and $\boldsymbol{\lambda}_r^d = \boldsymbol{\lambda}_{\neg nr}^d + \mathbf{z}_n^\top y_{nr}^d$ are the natural parameters of the Gaussian distribution.

The Accelerated Gibbs sampling algorithm iteratively samples the value of each element z_{nk} according to

$$p(z_{nk}=1|\{\mathbf{Y}^d\}_{d=1}^D,\mathbf{Z}_{\neg nk}) \propto \frac{m_{\neg n,k}}{N}\prod_{d=1}^{D}\prod_{r=1}^{S_d}\mathcal{N}(y_{nr}^d|\mathbf{z}_n\boldsymbol{\lambda}_{\neg nr}^d,\mathbf{z}_n\mathbf{P}_{\neg n}\mathbf{z}_n^\top+\sigma_y^2). \tag{5.12}$$

After having sampled all elements z_{nk} for the K_+ non-zero columns in $\mathbf{Z}$ for each data point n, the algorithm samples from a distribution (where the prior is a Poisson distribution with mean α/N) a number of new features necessary to explain that data point.

5.2.2 Posterior distribution over $\mathbf{Y}^d$

As previously described, in the 5-th step of Algorithm 3, we need to sample from the auxiliary Gaussian variables y_{nr}^d from the posterior distribution $p(y_{nr}^d|x_n^d,\mathbf{z}_n,\mathbf{b}^d)$. The posterior distribution y_{nr}^d for all the considered types of data are given given by:

1. For real-valued observation:

$$p(y_{n1}^d|x_n^d,\mathbf{z}_n,\mathbf{B}^d)=\delta(x_n^d) \tag{5.13}$$

2. For positive real-valued observations:

$$\begin{aligned}&p(y_{n1}^d|x_n^d,\mathbf{z}_n,\mathbf{B}^d)\\&=\mathcal{N}\left(y_{n1}^d\middle|\left(\frac{(\mathbf{z}_n\mathbf{b}_1^d)}{\sigma_y^2}+\frac{f^{-1}(x_n^d)}{\sigma_u^2}\right)\left(\frac{1}{\sigma_y^2}+\frac{1}{\sigma_u^2}\right)^{-1},\left(\frac{1}{\sigma_y^2}+\frac{1}{\sigma_u^2}\right)^{-1}\right).\end{aligned} \tag{5.14}$$

3. For categorical observations:

$$\begin{aligned}&p(y_{nr}^d|x_n^d=T,\mathbf{z}_n,\mathbf{B}^d)\\&=\begin{cases}\mathcal{N}(y_{nr}^d|\mathbf{z}_n\mathbf{b}_r^d,\sigma_y^2)\mathbb{I}(y_{nr}^d>\max_{j\neq r}(y_{nj}^d)) & \text{If} \quad r=T\\ \mathcal{N}(y_{nr}^d|\mathbf{z}_n\mathbf{b}_r^d,\sigma_y^2)\mathbb{I}(y_{nr}^d<y_{nT}^d) & \text{If} \quad r\neq T\end{cases}\end{aligned} \tag{5.15}$$

In words, if $x_n^d=T=r$ we sample y_{nr}^d from a Gaussian truncated by the left by $\max_{j\neq r}(y_{nj}^d)$ and, otherwise, we sample from a Gaussian truncated by the right by y_{nr}^d with $r=x_n^d$. Note that sampling from the variables y_{nr}^d corresponds to solve a multinomial probit regression problem. To achieve identifiability we assume, without loss of generality, that the regression function $f_{R_d}(\mathbf{z}_n)$ is identically zero, and therefore, we fix $b_{kR_d}^d=0$ for all k.

4. For ordinal observations:

$$p(y_{n1}^d | x_n^d = r, \mathbf{z}_n, \mathbf{B}^d) \sim \mathcal{N}(y_{n1}^d | \mathbf{z}_n \mathbf{b}_1^d, \sigma_y^2) \mathbb{I}(\theta_{r-1}^d < y_{n1}^d \leq \theta_r^d). \quad (5.16)$$

Note that in this case, we also need to sample the values for the thresholds θ_r^d with $r = 1, \ldots, R_d - 1$ as

$$\begin{aligned} p(\theta_r^d | y_{n1}^d) \sim & \mathcal{N}(\theta_r^d | 0, \sigma_\theta^2) \mathbb{I}(\theta_r^d > \max(\theta_{r-1}^d, \max_n(y_{n1}^d | x_n^d = r)) \\ & \times \mathbb{I}(\theta_r^d < \min(\theta_r^d, \min_n(y_{n1}^d | x_n^d = r+1)). \end{aligned} \quad (5.17)$$

In this case, sampling from the variables y_{n1}^d corresponds to solve an ordered probit regression problem, where the thresholds $\{\theta_r\}_{r=1}^{R_d}$ are unknown. Hence, to achieve identifiability we need to set the one of the thresholds, θ_1 in our case, to a fixed value.

5. For count observations:

$$p(y_{n1}^d | x_n^d, \mathbf{z}_n, \mathbf{B}^d) = \mathcal{N}(y_{n1}^d | \mathbf{z}_n \mathbf{b}_1^d, \sigma_y^2) \mathbb{I}(f^{-1}(x_n^d) \leq y_{n1}^d < f^{-1}(x_n^d + 1)), \quad (5.18)$$

where $f^{-1} : \Re_+ \to \Re$ is the inverse function of f, i.e., $f^{-1}(f(y)) = y$. Therefore, y_{n1}^d from a Gaussian truncated by the left by $f^{-1}(x_n^d)$ and by the right by $f^{-1}(x_n^d + 1)$.

Chapter 6

Analysis of Suicide Attempts

Every year, more than 34,000 suicides occur and over 370,000 individuals are treated for self-inflicted injuries in emergency rooms in the U.S., where suicide prevention is one of the top public health priorities [1]. The current strategies for suicide prevention have focused mainly on both detection and treatment of mental disorders [81], and on the treatment of the suicidal behaviors themselves [14]. However, despite prevention efforts including improvements in the treatment of depression, the lifetime prevalence of suicide attempts in the U.S. has remained unchanged over the past decade [38]. This suggests that there is a need to improve understanding of the risk factors for suicide attempts beyond psychiatric disorders, particularly in non-clinical populations.

According to the National Strategy for Suicide Prevention, an important first step in a public health approach to suicide prevention is to identify those at increased risk for suicide attempts [1]. Suicide attempts are, by far, the best predictor of completed suicide [56] and are also associated with major morbidity themselves [49]. The estimation of suicide attempt risk is a challenging and complex task, with multiple risk factors linked to increased risk. In the absence of reliable tools for identifying those at risk for suicide attempts, be they clinical or laboratory tests, risk detection still relays mainly on clinical variables. The adequacy of the current predictive models and screening methods has been questioned [56], and it has been suggested that the methods currently used for research on suicide risk factors and prediction models need revamping [44]. In the ongoing study, we aim at seeking the latent causes which lead to committing suicide as well as being able to detect those subjects that present higher risk of attempting suicide.

#	Source Code	Description
01	S4AQ4A17	Thought about committing suicide
02	S4AQ4A18	Felt like wanted to die
03	S4AQ17A	Stayed overnight in hospital because of depression
04	S4AQ17B	Went to emergency room for help because of depression
05	S4AQ4A19	Thought a lot about own death
06	S4AQ16	Went to counselor/therapist/doctor/other person for help to improve mood
07	S4AQ18	Doctor prescribed medicine/drug to improve mood/make you feel better
08	S4CQ15A	Stayed overnight in hospital because of dysthymia
09	S4AQ4A12	Felt worthless most of the time for 2+ weeks
10	S4CQ15B	Went to emergency room for help because of dysthymia
11	S4AQ52	Had arguments/friction with family, friends, people at work, or anyone else
12	S4AQ55	Spent more time than usual alone because didn't want to be around people
13	S4AQ21C	Used medicine/drug on own to improve low mood prior to last 12 months
14	S4AQ21A	Ever used medicine/drug on own to improve low mood/make self feel better
15	S4AQ20A	Ever drank alcohol to improve low mood/make self feel better
16	S4AQ20C	Drank alcohol to improve mood prior to last 12 months
17	S4AQ56	Couldn't do things usually did/wanted to do
18	S4AQ54	Had trouble doing things supposed to do -like working, doing schoolwork, etc.
19	S4AQ11	Any episode began after drinking heavily/more than usual
20	S4AQ15IR	Only/any episode prior to last 12 months began after drinking/drug use

Table 6.1: Enumeration of the 20 selected questions in the experiments, sorted in decreasing order according to their mutual information with the 'attempted suicide' question.

6.1 Experimental Setup

The NESARC includes a question about having attempted suicide as well as other related questions such as 'felt like wanted to die' and 'thought a lot about own death'. In this study, we build an unsupervised model with the 20 questions that present the highest mutual information with the suicide attempt question, which are shown in Table 6.1 together with

their code in the questionnaire. The 20 selected variables correspond to yes-or-no questions, which have four possible outcomes (i.e., $R = 4$): 'blank' (B), 'unknown' (U), 'yes' (Y) and 'no' (N). If a question is left blank the question was not asked.[1] If a question is said to be unknown either it was not answered or was unknown to the respondent.

We resort to the simplest IBP model for categorical observations proposed in Section 3, i.e., the model with only the binary matrix $\mathbf{Z}$ and weighting matrices $\mathbf{B}^d$ as latent variables. In order to sample from the IBP matrix, we make use of the Gibbs sampling algorithm combined with the Laplace approximation, detailed in Section 4.1.1, to compute the marginal likelihood. We initialize the sampler with an active feature, i.e., $K_+ = 1$, and set $z_{nk} = 1$ randomly with probability 0.5, and fixing $\alpha = 1$ and $\sigma_B^2 = 1$. Then, we run the Gibbs sampler over 500 randomly chosen subjects out of the 13,670 that have answered affirmatively to having had a period of low mood. In this study, we use another 9,500 subjects as test cases and have left the remaining samples for further validation.

6.2 Results

After running our inference algorithm, we obtain seven latent features. In Figure 6.1, we have plotted the posterior probability for each question when a single feature is active. In these plots, white means 0 and black 1, and each row sums up to one. Feature 1 is active for modeling the 'blank' and 'no' answers and, fundamentally, those who were not asked Questions 8 and 10. Feature 2 models the 'yes' and 'no' answers and favors affirmative responses to Questions 1, 2, 5, 9, 11, 12, 17 and 18, which indicates depression. Feature 3 models blank answers for most of the questions and negative responses to 1, 2, 5, 8 and 10, which are questions related to suicide. Feature 4 models the affirmative answers to 1, 2, 5, 9 and 11 and also have higher probability for unknowns in Questions 3, 4, 6 and 7. Feature 5 models the 'yes' answer to Questions 3, 4, 6, 7, 8, 10, 17 and 18, being ambivalent in Questions 1 and 2. Feature 6 favors 'blank' and 'no' answers in most questions. Feature 7 models answering affirmatively to Questions 15, 16, 19 and 20, which are related to alcohol abuse.

We show the percentage of respondents that answered positively to the

[1] In a questionnaire of this size some questions are not asked when a previous question was answered in a predetermined way to reduce the burden of taking the survey. For example, if a person has never had a period of low mood, the attempt suicide question is not asked.

suicide attempt questions in Table 6.2, independently for the 500 samples that were used to learn the IBP and the 9,500 hold-out samples, together with the total number of respondents. A dash indicates that the feature can be active or inactive. Table 6.2 is divided in three parts. The first part deals with each individual feature and the other two study some cases of interest. Throughout the database, the prevalence of suicide attempt is 7.83%. As expected, Features 2, 4, 5 and 7 favor suicide attempt risk, although Feature 5 only mildly, and Features 1, 3 and 6 decrease the probability of attempting suicide. From the above description of each feature, it is clear that having Features 4 or 7 active should increase the risk of attempting suicide, while having Features 3 and 1 active should cause the opposite effect.

Features 3 and 4 present the lowest and the highest risk of suicide, respectively, and they are studied together in the second part of Table 6.2, in which we can see that having Feature 3 and not having Feature 4 reduces this risk by an order of magnitude, and that combination is present in 70% of the population. The other combinations favor an increased rate of suicide attempts that goes from doubling ('11') to quadrupling ('00'), to a ten-fold increase ('01'), and the percentages of population with these features are, respectively, 21%, 6% and 3%.

In the final part of Table 6.2, we show combinations of features that significantly increase the suicide attempt rate for a reduced percentage of the population, as well as combinations of features that significantly decrease the suicide attempt rate for a large chunk of the population. These results are interesting as they can be used to discard significant portions of the population in suicide attempt studies and focus on the groups that present much higher risk. Hence, our IBP with discrete observations is being able to obtain features that describe the hidden structure of the NESARC database and makes it possible to pin-point the people that have a higher risk of attempting suicide.

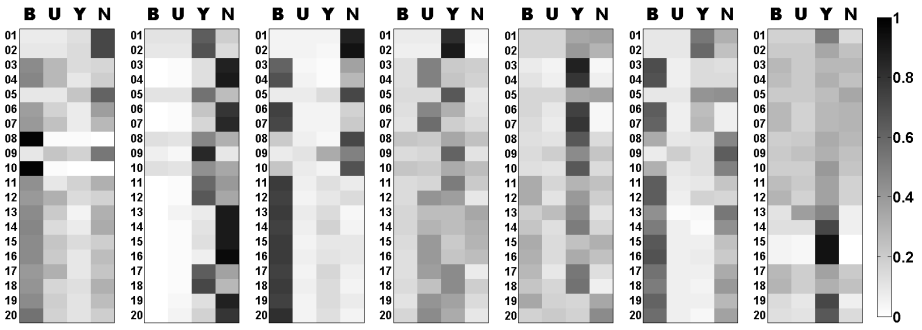


Figure 6.1: Probability of answering 'blank' (B), 'unknown' (U), 'yes' (Y) and 'no' (N) to each of the 20 selected questions, sorted as in Table 6.1. These probabilities have been obtained with the posterior mean weights $\mathbf{B}^d_{\text{MAP}}$, when only one of the seven latent features (sorted from left to right to match the order in Table 6.2) is active.

Hidden features							Suicide attempt probability		Number of cases	
							Train	Hold-out	Train	Hold-out
1	-	-	-	-	-	-	6.74%	5.55%	430	8072
-	1	-	-	-	-	-	10.56%	11.16%	322	6083
-	-	1	-	-	-	-	**3.72**%	**4.60**%	457	8632
-	-	-	1	-	-	-	**25.23**%	**22.25**%	111	2355
-	-	-	-	1	-	-	8.64%	9.69%	301	5782
-	-	-	-	-	1	-	6.90%	7.18%	464	8928
-	-	-	-	-	-	1	14.29%	14.18%	91	1664
-	-	0	0	-	-	-	30.77%	28.55%	26	571
-	-	0	1	-	-	-	**82.35**%	**61.95**%	17	297
-	-	1	0	-	-	-	**0.83**%	**0.87**%	363	6574
-	-	1	1	-	-	-	14.89%	16.52%	94	2058
-	-	0	1	-	-	1	100.00%	69.41%	4	85
0	-	0	1	-	-	-	80.00%	66.10%	5	118
1	-	1	0	-	1	0	0.00%	0.25%	252	4739
-	-	1	0	-	-	0	0.33%	0.63%	299	5543
1	-	1	0	-	-	-	**0.32**%	**0.41**%	317	**5807**

Table 6.2: Probabilities of attempting suicide for different values of the latent feature vector, together with the number of subjects possessing those values. The symbol '-' denotes either 0 or 1. The 'train ensemble' columns contain the results for the 500 data points used to obtain the model, whereas the 'hold-out ensemble' columns contain the results for the remaining subjects.

Chapter 7

Analysis of Psychiatric Disorders

7.1 Comorbidity Analysis

Health care increasingly needs to address the management of individuals with multiple coexisting diseases, who are now the norm, rather than the exception. In the United States, about 80% of Medicare spending is devoted to patients with four or more chronic conditions, with costs growing as the number of chronic conditions increases [96]. This explains the growing interest of researchers in the impact of comorbidity on a range of outcomes, such as mortality, health-related quality of life, functioning, and quality of health care. However, attempts to study the impact of comorbidity are complicated due to the lack of consensus about how to define and measure it [85].

Comorbidity becomes particularly relevant in psychiatry, where clinical experience and several studies suggest that the relation among the psychiatric disorders may have etiological and treatment implications. Several studies have focused on the search of the underlying interrelationships among psychiatric disorders, which can be useful to analyze the structure of the diagnostic classification system, and guide treatment approaches for each disorder [7]. In [43], the authors found that 10 psychiatric disorders (available in the National Comorbidity Survey) can be explained by only two correlated factors, one corresponding to internalizing disorders and the other to externalizing disorders. The existence of the internalizing and the externalizing factors was also confirmed by [42]. More recently, the authors in [7] have used factor analysis to find the latent

feature structure under 20 common psychiatric disorders, drawing on data from the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). In particular, the authors found that three correlated factors, one related to externalizing, and the other two to internalizing disorders, characterized well the underlying structure of these 20 diagnoses. From a statistical point of view, the main limitation of this study lies on the use of factor analysis, which assumes that the number of factors is known and that the observations are Gaussian distributed. However, the latter assumption does not fit the observed data, since they are discrete in nature.

In the present study, our objective is to provide an alternative to the factor analysis approach used by [7] with the IBP models in Chapter 3. In particular, we build an unsupervised model taking the 20 disorders used by [7] as input data, drawn from the NESARC data. These disorders include substance use disorders (alcohol abuse and dependence, drug abuse and dependence and nicotine dependence), mood disorders (major depressive disorder (MDD), bipolar disorder and dysthymia), anxiety disorders (panic disorder, social anxiety disorder (SAD), specific phobia and generalized anxiety disorder (GAD)), pathological gambling (PG) and seven personality disorders (avoidant, dependent, obsessive-compulsive (OC), paranoid, schizoid, histrionic and antisocial personality disorders (PDs)).

The main goal of this study is to find out and analyze the latent relations among the 20 psychiatric disorders. Specifically, we aim at finding comorbidity patterns in the database, allowing us to seek hidden causes and to provide a tool for detecting those subjects with higher risk of suffering from these disorders.

7.1.1 Experimental Setup

Based on information collected in the first wave of the NESARC, a set of pre-established and reliable diagnostic algorithms were applied to each subject to determine the presence or absence of 20 psychiatric disorders [7]. In this study, we use these diagnoses as input data to the full IBP model in Figure 3.3, i.e., the IBP model with bias term and severity matrix. We assume a multinomial-probit likelihood model with two categories (i.e., positive and negative diagnoses), and resort to the MH based sampler, detailed in Section 4.1.3, to jointly infer the IBP matrix $\mathbf{Z}$ and the severity matrix $\mathbf{W}$, being weighting factors in $\mathbf{B}^d$ integrated out using the EP approximation.

For the following experimental results, we set $\alpha = 1$, $\sigma_B^2 = 1$, $\gamma_1 = 2$ and $\gamma_2 = 1$ and run our inference algorithm. In order to speed up the

inference procedure, we do not sample the rows of $\mathbf{W}$ corresponding to those subjects who suffer from at most one out of the 20 disorders, but instead fix these latent features to zero. The idea is that the $\mathbf{b}_0^d$ terms must capture the general population, and we use the active components of the matrix $\mathbf{W}$ to characterize the disorders. Besides speeding up the algorithm, this modification ensures that the active latent features increase the probability of suffering from the disorders and can be interpreted as latent disorders, which helps the psychiatrists understand the obtained results. If we had no bias term and did not force the subjects in the general population to be explained by the bias term, the latent variables would not be easy to interpret because the general population would be described by a combination of latent factors.

7.1.2 Results

Similar to the previous study in [7], we find that we need three latent features to describe the data. In Table 7.1, we show the empirical probability of possessing each of the inferred latent feature, i.e., the number of subjects in the database that possess each latent feature divided by the total number of subjects. Additionally, we plot in Figure 7.1 the approximate posterior probability of suffering from each of the considered disorders when only one of the latent features is active (assuming severity factors equal to one), and when none of them is active. As expected, for those subjects without any active latent feature, the probability of having any disorder is below the baseline level (defined as the empirical probability of suffering from each disorder in the full database).

We can interpret each of the obtained latent features from the analysis of Figure 7.1. Feature 1 (pattern [100]) increases the probability of having all disorders, except alcohol abuse, and thus seems to represent a general psychopathology factor, although it may particularly increase the risk of personality disorders (disorders from 14 to 20). Feature 2 (pattern [010]) models substance use disorders and antisocial personality disorder, which is consistent with the externalizing factor identified in previous studies of the structure of psychiatric disorders [43, 37, 88, 7]. Feature 3 (pattern [001]) models mood or anxiety disorders, and thus seems to represent the internalizing factor also identified in previous studies. Note that the probability of bipolar disorder presents a significantly different behavior, since major depression (MDD) and dysthymia are mutually exclusive with bipolar disorder.

In addition to the hidden relation among the disorders, our model also

provides an individual-specific severity term that can be interpreted as our belief in the subject suffering a latent disorder. We find that more than 80% of the subjects with active features have a severity factor above 0.5 and around 50% of them have a severity value greater that 0.75. The histograms for w_{n1}, w_{n2} and w_{n3} are shown in Figure 7.6. In order to examine the effect of the severity, we plot in Figure 7.2 the posterior probability of suffering from each of the disorders when only Feature 1 is active, for any value of the severity w_{n1}. (Similar plots, for Features 2 and 3, are provided in Figures 7.3 and 7.4, respectively.) When the severity reaches 0 (depicted in black), Feature 1 turns inactive and, therefore, the corresponding probabilities coincide with the green line in Figure 7.1 (pattern [000]). As the severity approaches 1 (depicted in red), the corresponding probabilities coincide with the red line in Figure 7.1 (pattern [100]). The solid line in Figure 7.2 represents the empirical probability of suffering from each disorder, obtained for those subjects who only have Feature 1 active. We can see, that although the probability of suffering from each disorder becomes higher when the inferred severity value increases, each disorder is affected differently by the value of the severity factor. For instance, the probability of suffering from OCPD goes from 0.04 in the general population to 0.8 for the subjects with a severity factor for Feature 1 near to one, while the probability for alcohol abuse only changes from 0.04 to 0.05.

To further analyze the impact of severity, we depict in Figure 7.5 the distribution of the number of disorders for those subjects whose inferred severity is comprised between the numbers shown in the horizontal axis. As expected, as the inferred severity increases, so does the number of disorders that a subject suffers. Figure 7.5a shows that Feature 1 (general psychopathology factor) is the feature with highest impact on the average number of disorders. However, when we only consider a subset of the disorders (Figures 7.5b and 7.5c), Features 2 and 3 become more relevant. These subsets have been chosen to match the externalizing and internalizing factors, respectively.

From the analysis of the figures, we can conclude that the probability of appearance of a disorder changes significantly when the value of the severity associated to that group of disorders changes. We also find that most of the subjects with active latent features suffer from three or more disorders and, in general, most of the disorders that a subject suffers belong to the group of disorders modeled by the same latent feature. Therefore, a subject with Feature 2 (Feature 3) active has a higher probability of suffering simultaneously from several externalizing (internalizing) disorders. Finally, we can understand the importance of the severity factors in the model,

because they allow explaining the comorbidity among the disorders and also understanding the stress each subject suffers. The model without the severity factors cannot distinguish between the different subjects that have the same active latent features.

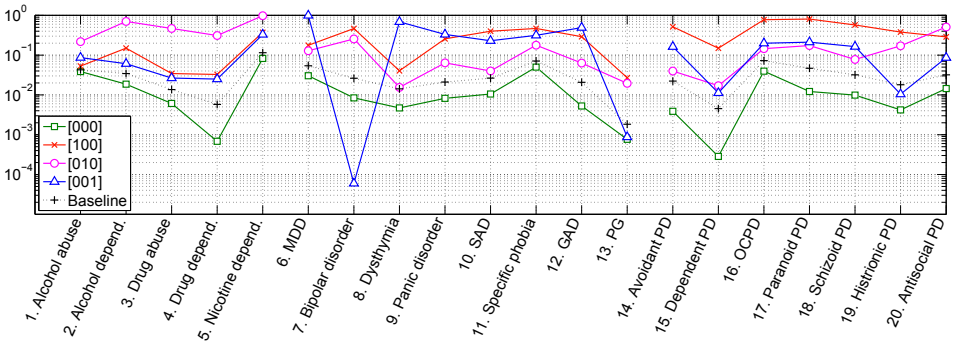


Figure 7.1: Probabilities of suffering from the 20 considered disorders for the latent feature vectors $\mathbf{w}_n$ shown in the legend. These probabilities have been obtained using the mean of the approximate posterior of the matrices $\mathbf{B}^d$.

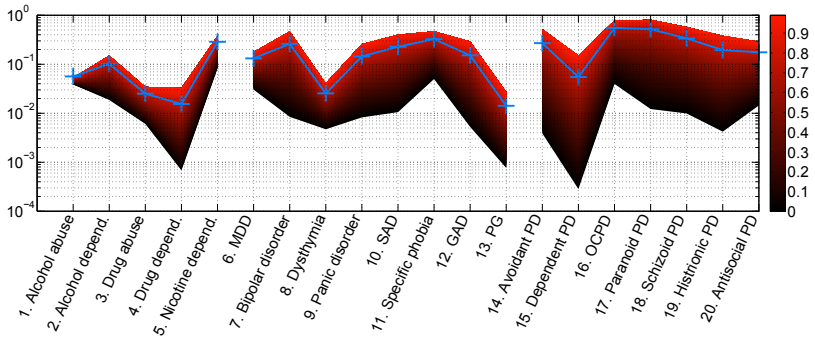


Figure 7.2: Probabilities of suffering from the 20 considered disorders when only Feature 1 is active, for any value of the severity w_{n1} (shown in the bar on the right). These probabilities have been obtained using the mean of the approximate posterior of the matrices $\mathbf{B}^d$. The solid line represents the empirical probabilities, obtained for those subjects who only have Feature 1 active.

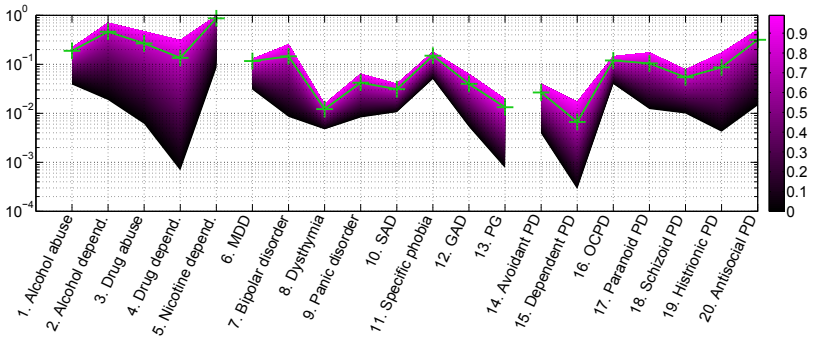


Figure 7.3: Probabilities of suffering from the 20 considered disorders when only Feature 2 is active, for any value of the severity w_{n2} (shown in the bar on the right). These probabilities have been obtained using the mean of the approximate posterior of the matrices $\mathbf{B}^d$. The solid line represents the empirical probabilities, obtained for those subjects who only have Feature 2 active.

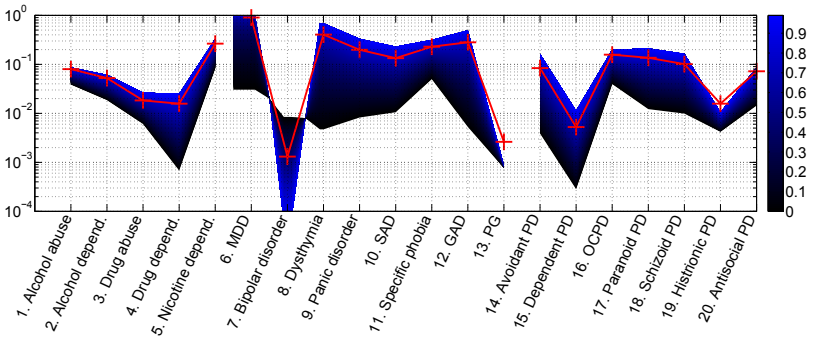
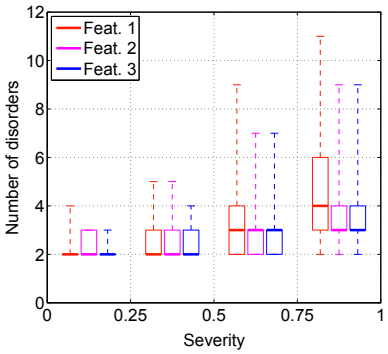


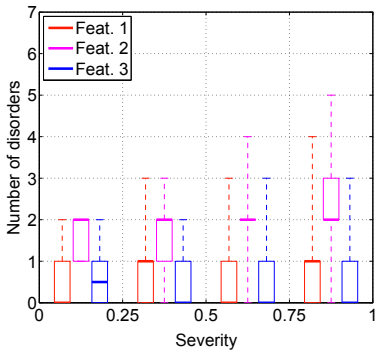
Figure 7.4: Probabilities of suffering from the 20 considered disorders when only Feature 3 is active, for any value of the severity w_{n3} (shown in the bar on the right). These probabilities have been obtained using the mean of the approximate posterior of the matrices $\mathbf{B}^d$. The solid line represents the empirical probabilities, obtained for those subjects who only have Feature 3 active.

Active Feature	Feature 1	Feature 2	Feature 3
Empirical Prob.	0.0594	0.0239	0.0201

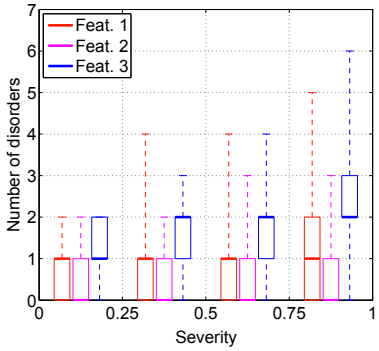
Table 7.1: Empirical probabilities of possessing at least one latent feature, extracted directly from the inferred IBP matrix $\mathbf{Z}$.



(a) 20 disorders.



(b) Externalizing disorders (disorders 1 to 5 in Fig. 7.1 and antisocial PD).



(c) Internalizing disorders (disorders 6 and 8-12 in Fig. 7.1).

Figure 7.5: Distribution of the number of disorders, for those subjects who only have active one latent feature (shown in the legend), whose inferred severity is comprised between the numbers shown in the horizontal axis. The thick line corresponds to the median, the edges of the box are the 25th and 75th percentiles, and the whiskers represents the most extreme values.

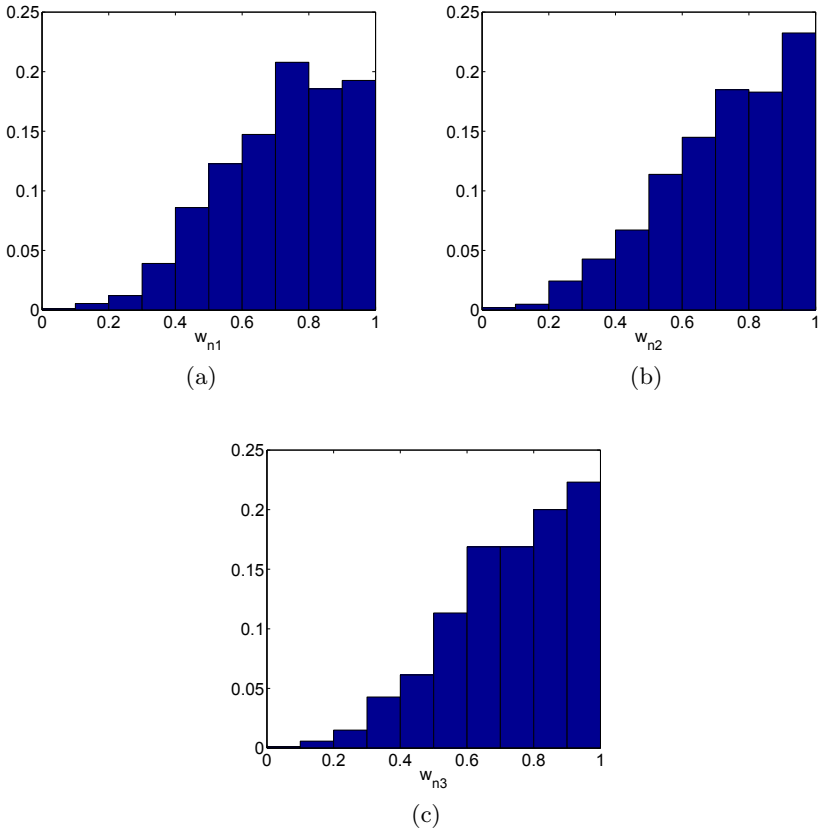


Figure 7.6: Normalized histograms of w_{n1}, w_{n2} and w_{n3} (assuming that $z_{n1} = 1$, $z_{n2} = 1$ and $z_{n3} = 1$, respectively).

7.2 Impact of Social Background

Several studies have analyzed the impact of social background in the development of mental disorders. These studies usually focus on the relation between a mental disorder and a specific aspect of the social background of the subjects. Some examples in this area study the relation between depression and sex [92, 39], relation between common mental disorders and poverty or social class [91, 17, 31], etc. However, up to our knowledge, there is a lack of work in the study of the impact the social background in the suffering of comorbid disorders.

In the previous section, we found that the 20 psychiatric disorders under study can be divided into three groups, namely internalizing, externalizing

and personality disorders. We also found that comorbid disorders tend to belong to the same group. In this section, we aim at studying how the social background of the subjects (such as the age, sex, etc.) shows up in the comorbidity patterns studied above. To this end, we include in our experiments the responses to some of the questions in Section 1 of the NESARC, which collects information about the social background of the participants. Specifically, we incorporate the following information: sex, age, census region, imputed race/ethnicity, marital status, highest grade or years of school completed, and the body mass index (BMI).

7.2.1 Experimental Setup

In addition to the diagnoses of 20 psychiatric disorders in the previous section, we include one by one the background questions as input data to the IBP model. In this study, we make use of the model and inference algorithm introduced in Chapter 5 because they allow us to deal with all the considered questions. In Table 7.2, we summarize the considered questions and how we introduce them into our model as input variables.

For the following experimental results, we independently run the inference algorithm in Section 5.2 for each question with $\alpha = 5$, $\sigma_B^2 = 1$, $\sigma_y^2 = 1$, $\sigma_\theta^2 = 1$, and consider for the real positive and the count data the following transformation that maps from the real numbers to the real positive numbers: $f(x) = ax^2$, where a is a hyperparameter. In this study, we do not sample the rows of $\mathbf{Z}$ corresponding to those subjects who do not suffer from any of the 20 disorders, but instead fix these latent features to zero. The idea is that the $\mathbf{b}_0^d$ terms must capture the general population, and we use the active components of the matrix $\mathbf{Z}$ to characterize the disorders.

Description	Type of variable
Sex	Categorical with 2 categories
Age	Count data
Census region	Categorical with 4 categories
Race/ethnicity	Categorical with 5 categories
Marital status	Categorical with 6 categories
Highest grade or years of school completed	Ordinal with 14 categories
BMI	Positive real

Table 7.2: Enumeration of the 8 selected questions related to the social background of the subjects.

7.2.2 Results

1. Sex. We model the gender information of the participants in the NESARC as a categorical variable with two categories: {'male', 'female'}, being percentage of males in the NESARC around 43%. After running our inference algorithm with the diagnoses of the 20 disorders and the sex of the subjects as input data, we obtain three latent features. In Table 7.3, we show the empirical probability of possessing each of the inferred latent features, i.e., the number of subjects in the database that possess each latent feature divided by the total number of subjects. In Figure 7.7a, we show the probability of meeting each diagnostic criteria for the latent feature vectors $\mathbf{z}_n$ shown in the legend and in the database (baseline). Note that the obtained latent features are similar to the ones in Figure 7.1, i.e., Feature 1 (pattern [100]) mainly models the seven PDs, Feature 2 (pattern [010]) models the alcohol and drug abuse disorders and the antisocial PD, and Feature 3 (pattern [001]) models the anxiety and mood disorders. Additionally, in Figure 7.7b, we show the probability of being male and female for the latent feature vectors $\mathbf{z}_n$ shown in the legend and the probability of being male and female in the database (baseline). In Figure 7.7b, we observe that having not active features (pattern [000]), which model people that so not suffer from any disorder, increases the probability of being male with respect to the baseline probability, and therefore, it indicates that females tend to suffer in a higher extent from psychiatric disorders. Additionally, we observe that being male increases the probability of Feature 1 (pattern [100]), while being female increases the probability of Feature 3 (pattern [001]). Hence, from the analysis of Figure 7.7b, we can conclude that, while women frequently suffer from mood and anxiety disorders, PDs more often appear in men.

Active Feature	Feature 1	Feature 2	Feature 3
Empirical Prob.	0.0341	0.0470	0.0460

Table 7.3: Sex. Empirical probabilities of possessing at least one latent feature, extracted directly from the inferred IBP matrix $\mathbf{Z}$.

2. Age. Now, we focus on the age of the participants, which we model as count data[1]. The numerical probability distribution over the age based on the data is shown (and denoted by 'baseline') in Figure 7.8b. After running our inference algorithm with the diagnoses of the 20 disorders and the age

[1]We set the hyperparameter a in the transformation $f(x) = ax^2$ to 1.

of the subjects as input data, we obtain three latent features. In Table 7.4, we show the empirical probability of possessing each inferred latent feature. Figure 7.8a shows the probability of meeting each diagnostic criteria for the latent feature vectors $\mathbf{z}_n$ shown in the legend and in the database (baseline). In addition to the baseline probability distribution, we plot in Figure 7.8b the inferred probability distributions over the age when none or one of the latent variables is active. In Figure 7.8b, we observe that introducing the age of the participants as an input variable has change (with respect to the features in Figure 7.1) the inferred latent features. In particular, we observe that the obtained latent features mainly differ in the probability of suffering from personality disorders (i.e., disorders from 14 to 20), being the probability of suffering from disorders 1 to 13 similar for the three plotted latent feature vectors. In this figure, we observe that the vector $\mathbf{z}_n$ with no active latent features (pattern [000]) is trying to capture the mean of the age in the database (which coincides with middle-aged subjects, i.e., $30-50$ years old). Moreover, we observe that the subjects with the highest probability of suffering from personality disorders (pattern [100]) are likely to be middle-aged, followed in a decreasing order by young adults (pattern [010]) and elderly people (pattern [001]). Additionally, if we focus on the differences among the three features in disorders from 1 to 13, we also observe that, while young and elderly people tend to suffer from depression, middle-aged people tend to suffer from the bipolar disorder. Hence, based on Figure 7.8, we can conclude that the bipolar disorder and the seven personality disorders tend to show up in a higher extent in the mature age, while young and elderly people tend to suffer from depression.

Active Feature	Feature 1	Feature 2	Feature 3
Empirical Prob.	0.0332	0.0550	0.0569

Table 7.4: Age. Empirical probabilities of possessing at least one latent feature, extracted directly from the inferred IBP matrix $\mathbf{Z}$.

3. Census Region. We model the census region information of the participants in the NESARC as a categorical variable with four categories, {'northeast', 'midwest', 'south', 'east'}. After running our inference algorithm with the diagnoses of the 20 disorders and the census region information of the subjects as input data, we obtain three latent features. In Table 7.3, we show the empirical probability of possessing each of the inferred latent features, i.e., the number of subjects in the database that possess each latent feature divided by the total number of subjects. In

Figure 7.7a, we show the probability of meeting each diagnostic criteria for the latent feature vectors $\mathbf{z}_n$ shown in the legend and in the database (baseline). Note that the obtained latent features are similar to the ones in Figure 7.1, i.e., Feature 1 (pattern [100]) mainly models all the PDs, Feature 2 (pattern [010]) models the alcohol and drug abuse disorders and the antisocial PD, and Feature 3 (pattern [001]) models the anxiety and mood disorders. Additionally, in Figure 7.7b, we show the probability of belonging to each region for the latent feature vectors $\mathbf{z}_n$ shown in the legend and in the database (baseline). In Figure 7.7b, we can observe that the inferred probabilities are in general similar for the four considered latent vectors $\mathbf{z}_n$ and the baseline, except for Feature 1, which models the PDs and slightly increases the probability of living in the northeast and decreases the probability of living in the west of the U.S. Hence, since we do not appreciate significant statistical differences, we can conclude that the location of the subjects does not appear as an influential variable in the suffering of the obtained latent psychiatric disorders.

Active Feature	Feature 1	Feature 2	Feature 3
Empirical Prob.	0.0335	0.0385	0.0440

Table 7.5: Census Region. Empirical probabilities of possessing at least one latent feature, extracted directly from the inferred IBP matrix $\mathbf{Z}$.

4. Race/Ethnicity. We model the race information of the participants in the NESARC as a categorical variable with five categories, {'White, not Hispanic or Latino', 'Black, not Hispanic or Latino', 'American Indian/Alaska native, not Hispanic or Latino', 'Asian/Native Hawaiian/Pacific Islander, not Hispanic or Latino', 'Hispanic or Latino'}. After running our inference algorithm with the diagnoses of the 20 disorders and the race of the subjects as input data, we obtain three latent features. In Table 7.6, we show the empirical probability of possessing each of the inferred latent feature, i.e., the number of subjects in the database that possess each latent feature divided by the total number of subjects. In Figure 7.10a, we show the probability of meeting each diagnostic criteria for the latent feature vectors $\mathbf{z}_n$ shown in the legend and in the database (baseline). Note that the obtained latent features are similar to the ones in Figure 7.1, i.e., Feature 1 (pattern [100]) models the PDs, Feature 2 (pattern [010]) models the alcohol and drug abuse disorders and the antisocial PD, and Feature 3 (pattern [001]) models the anxiety and mood disorders. Additionally, in Figure 7.10b, we show the probability of belonging to each ethnic group

for the latent feature vectors $\mathbf{z}_n$ shown in the legend and in the database (baseline). In this figure, we observe that all the probabilities, except the pattern [100] for American Indian and Alaska natives (which can either be due to a poor estimation of the probability of the less common race or mean that American Indian and Alaska natives suffer in a less extent from PDs), are close to the baseline and, therefore, we can conclude that the race of the subjects does not influence the presence or absence of any of the three latent psychiatric disorders (internalizing, externalizing or personality disorders).

Active Feature	Feature 1	Feature 2	Feature 3
Empirical Prob.	0.0356	0.0248	0.0533

Table 7.6: Race. Empirical probabilities of possessing at least one latent feature, extracted directly from the inferred IBP matrix $\mathbf{Z}$.

5. Marital Status. We now perform a similar analysis with the marital status of the subjects, which we model as a categorical variable with six categories, {'Married', 'Living with someone as if married (not currently married or separated from another person)', 'Widowed', 'Divorced', 'Separated', 'Never Married' }. After running our inference algorithm with the diagnoses of the 20 disorders and the marital status of the subjects as input data, we obtain three latent features. In Table 7.7, we show the empirical probability of possessing each of the inferred latent features. In Figure 7.11a, we show the probability of meeting each diagnostic criteria for the latent feature vectors $\mathbf{z}_n$ shown in the legend and in the database (baseline). Additionally, in Figure 7.11b, we show the probability of each marital status for the latent feature vectors $\mathbf{z}_n$ shown in the legend and in the database (baseline). Since the probabilities under the four patterns are similar to the probabilities in the baseline (except the pattern [001] for the 'living with someone' category), we can conclude from Figure 7.11b that the marital status of the subjects does not influence the presence or absence of any of the three latent psychiatric disorders (internalizing, externalizing or personality disorders). The increase in the probability of being 'Living with someone as if married (not currently married or separated from another person)' under the pattern [001], seems to indicate that these subjects tend to suffer in a higher extent from mood or anxiety disorders.

6. Highest Grade of School Completed. We now include in our analysis the information about the grade of studies of the subjects, which we model as an ordinal variable with the following fourteen categories, {'No formal schooling', 'completed grade K, 1 or 2', 'completed grade 3 or 4', 'completed

Active Feature	Feature 1	Feature 2	Feature 3
Empirical Prob.	0.0362	0.0433	0.0404

Table 7.7: Marital Status. Empirical probabilities of possessing at least one latent feature, extracted directly from the inferred IBP matrix $\mathbf{Z}$.

Active Feature	Feature 1	Feature 2	Feature 3
Empirical Prob.	0.0341	0.0379	0.0469

Table 7.8: School. Empirical probabilities of possessing at least one latent feature, extracted directly from the inferred IBP matrix $\mathbf{Z}$.

grade 5 or 6', 'completed grade 7', 'completed grade 8', 'some high school (grades 9-11)', 'completed high school', 'graduate equivalency degree (GED)', 'some college (no degree)', 'completed associate or other technical 2-year degree', 'completed college (bachelor s degree)', 'some graduate or professional studies (completed bachelor's degree but not graduate degree)', 'completed graduate or professional degree (master's degree or higher)'}. After running our inference algorithm with the diagnoses of the 20 disorders and the level of studies of the subjects as input data, we obtain three latent features. In Table 7.8, we show the empirical probability of possessing each of the inferred latent features, i.e., the number of subjects in the database that possess each latent feature divided by the total number of subjects. In Figure 7.12a, we show the probability of meeting each diagnostic criteria for the latent feature vectors $\mathbf{z}_n$ shown in the legend and in the database (baseline). This figure shows that, although the three obtained features increase the probability of suffering from all the disorders, Feature 1 (pattern [010]) mainly increases the probability of suffering from disorders 13 to 20, i.e., personality disorders, Feature 2 (pattern [010]) mainly increases the probability of suffering from disorders 1 to 5, i.e., drug and alcohol abuse disorders (externalizing factor), and Feature 3 (pattern [001]) mainly increases the probability of suffering from disorders 6 to 13, i.e., mood disorders (internalizing factor). Additionally, in Figure 7.12b, we show the probability of having each level of studies for the latent feature vectors $\mathbf{z}_n$ shown in the legend and in the database (baseline). In this figure, we observe that pattern [000] which models the people that do not suffer from any disorder, decreases the probability of having a lower level or grade of studies, increasing, in turn, the probability of suffering from any of the latent disorders.

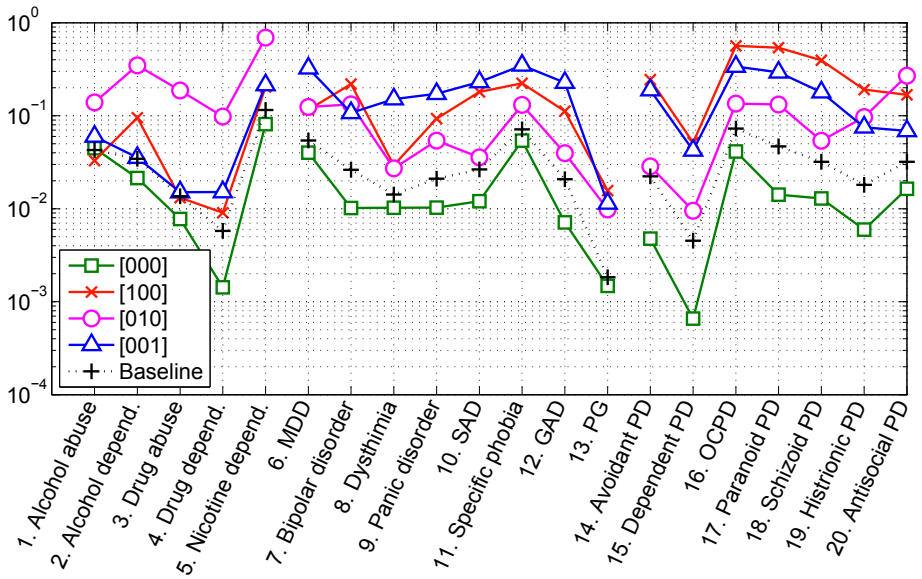
7. Body Mass Index. Finally, since the NESARC collect the weight and heigh of the participants, we have computed the BMI fo each subject as $BMI = \frac{mass(lb)}{(height(in))^2} \times 703$. Figure 7.13b (dashed line) shows the estimated probability density[2] given by the data in the NESARC. Note that a BMI bellow 18.5 is classified as underweight, a BMI between 18.5 and 25 corresponds to normal (healthy weight), between 25 and 30 is classified as overweight, and larger than 30 is classified as obesity. After running our inference algorithm with the diagnoses of the 20 disorders and the BMI as input data (being the BMI modeled as a positive real-valued variable[3]), we obtain four latent features. In Table 7.9, we show the empirical probability of possessing each of the inferred latent feature, i.e., the number of subjects in the database that possess each latent feature divided by the total number of subjects. In Figure 7.13a, we show the probability of meeting each diagnostic criteria for the latent feature vectors $\mathbf{z}_n$ shown in the legend and in the database (baseline). In this figure, we observe that Feature 1 (pattern [1000]) models the seven PDs, Feature 2 (pattern [0100]) models the alcohol and drug abuse disorders and the antisocial PD, Feature 3 (pattern [0010]) models the anxiety and mood disorders , and Feature 4 (pattern [0001]) is similar to Feature 1 but it presents higher probability of suffering from alcohol abuse disorder, mayor depression disorder (MDD) and dysthimia. Additionally, Figure 7.13b shows the (estimated) baseline probability density and the probability density over the BMI for the latent feature vectors $\mathbf{z}_n$ shown in the legend. Note that the pattern [0001] is trying to capture the probability over the BMI in the database (baseline), matching the mean of the BMI in the database. In Figure 7.8b, we observe that the suffering of the seven PDs do not depend on the BMI, since Features 1 and 4 model the seven PDs and cover all the possible values for the BMI. People with Feature 4 present a lower BMI and tend to suffer in a higher extent from alcohol abuse, mayor depression and dysthimia disorders than people with Feature 1. Additionally, note that people that suffer from alcohol and drug use disorders tend to present a lower BMI than general population (baseline) while people that suffer from mood and anxiety disorders present a larger BMI than the baseline.

[2]The estimated probability density is computed using the Matlab function 'ksdensity', in which the estimate is based on a normal kernel function, and is evaluated at 100 equally spaced points that cover the range of the data.

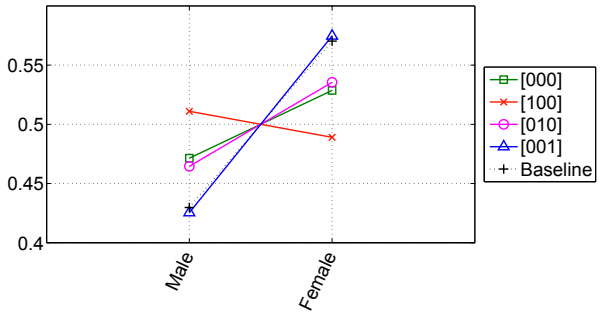
[3]We set the hyperparameter a in the transformation $f(x) = ax^2$ to 0.25.

Active Feature	Feature 1	Feature 2	Feature 3	Feature 4
Empirical Prob.	0.0420	0.0404	0.0227	0.0223

Table 7.9: BMI. Empirical probabilities of possessing at least one latent feature, extracted directly from the inferred IBP matrix $\mathbf{Z}$.

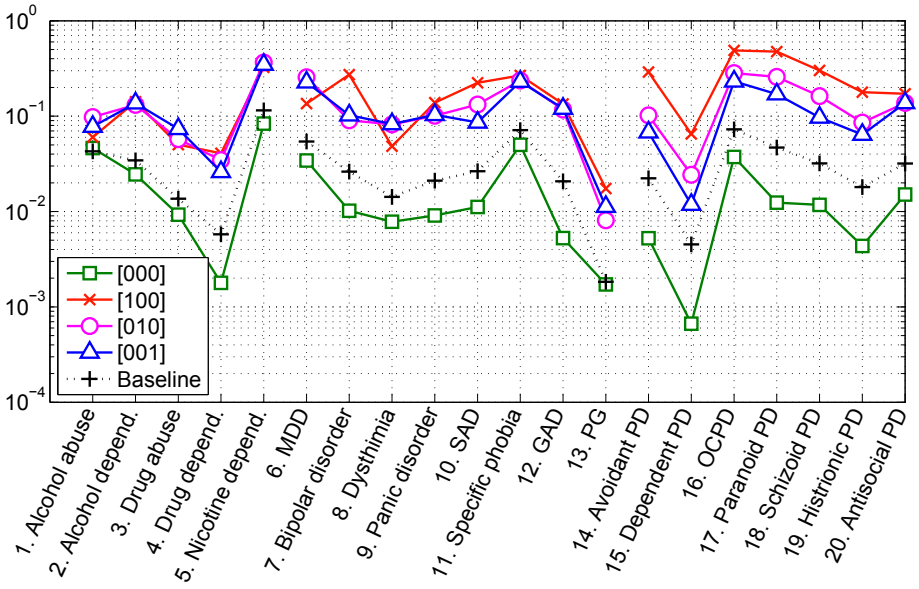


(a)

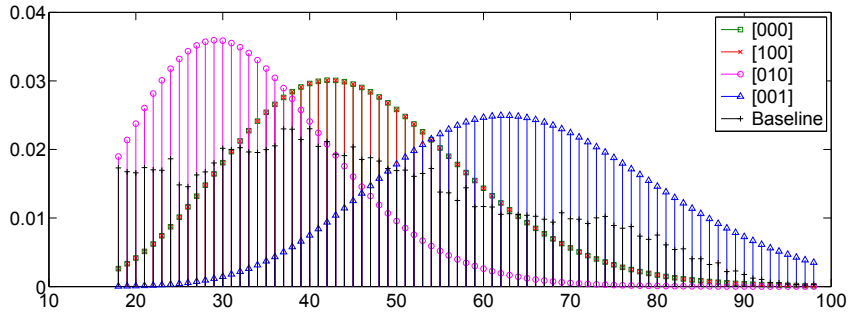


(b)

Figure 7.7: Sex. (a) Probabilities of suffering from the 20 considered disorders and (b) probability of belonging to each category for the latent feature vectors $\mathbf{z}_n$ shown in the legend and for the baseline.

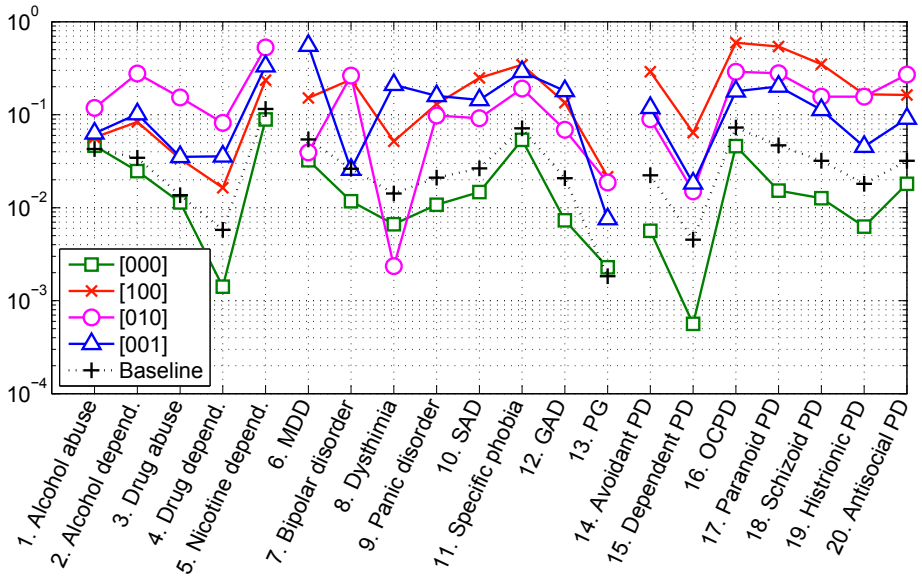


(a)

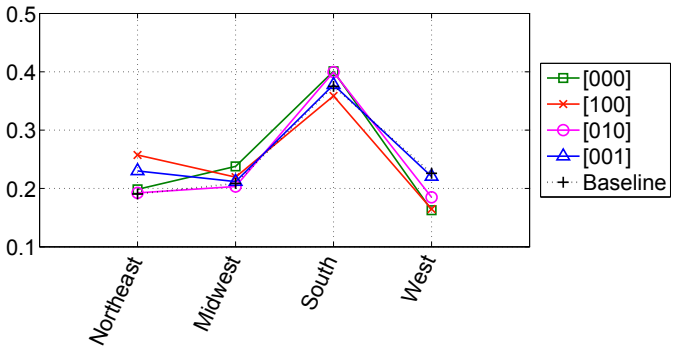


(b)

Figure 7.8: Age. (a) Probabilities of suffering from the 20 considered disorders for the latent feature vectors $\mathbf{z}_n$ shown in the legend and (b) inferred probability distribution for the latent feature vectors $\mathbf{z}_n$ shown in the legend and baseline probability distribution.

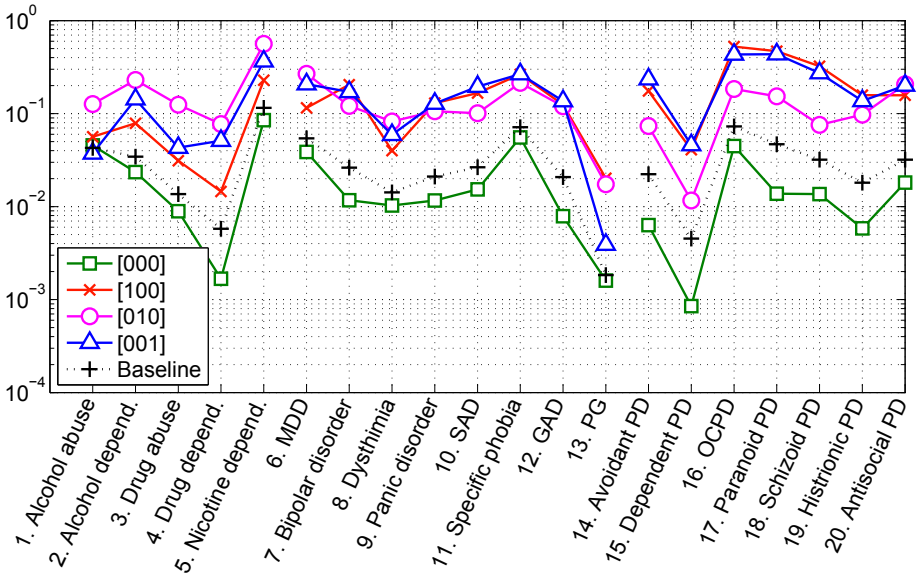


(a)

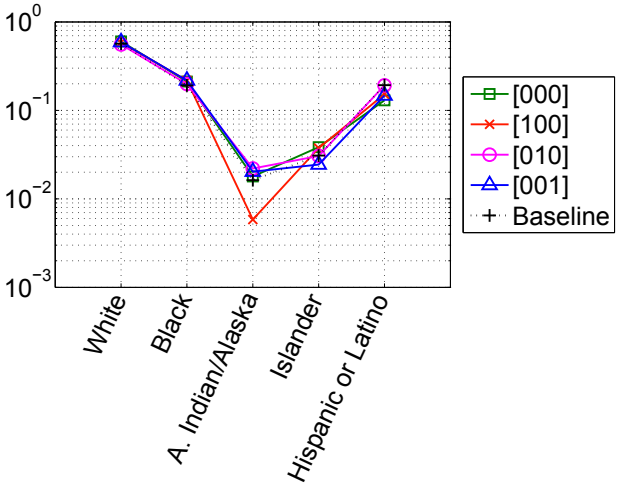


(b)

Figure 7.9: Census Region. (a) Probabilities of suffering from the 20 considered disorders and (b) probability of belonging to each category for the latent feature vectors $\mathbf{z}_n$ shown in the legend and for the baseline.

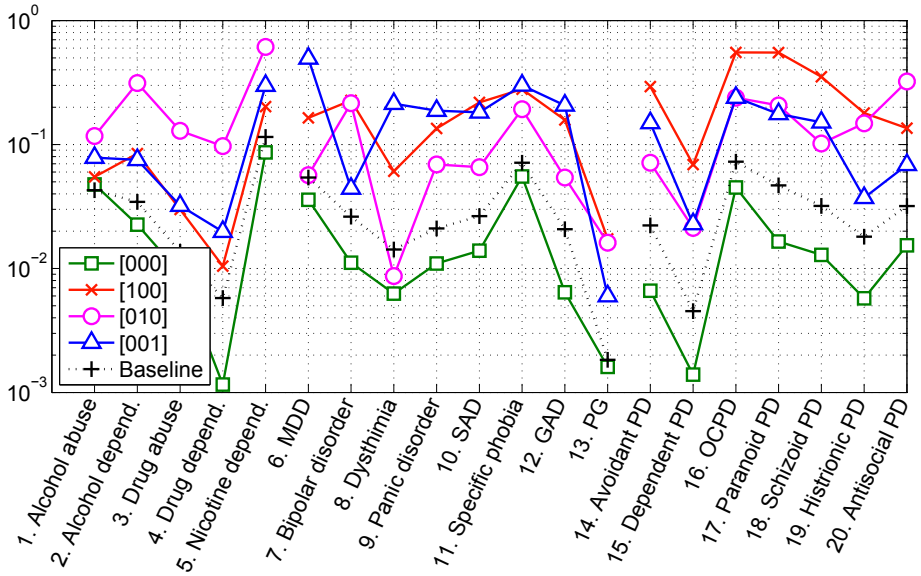


(a)

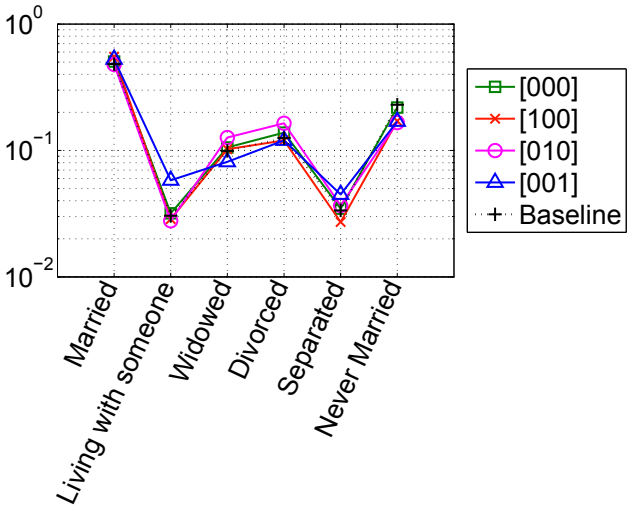


(b)

Figure 7.10: Race. (a) Probabilities of suffering from the 20 considered disorders and (b) probability of belonging to each category for the latent feature vectors $\mathbf{z}_n$ shown in the legend and for the baseline.

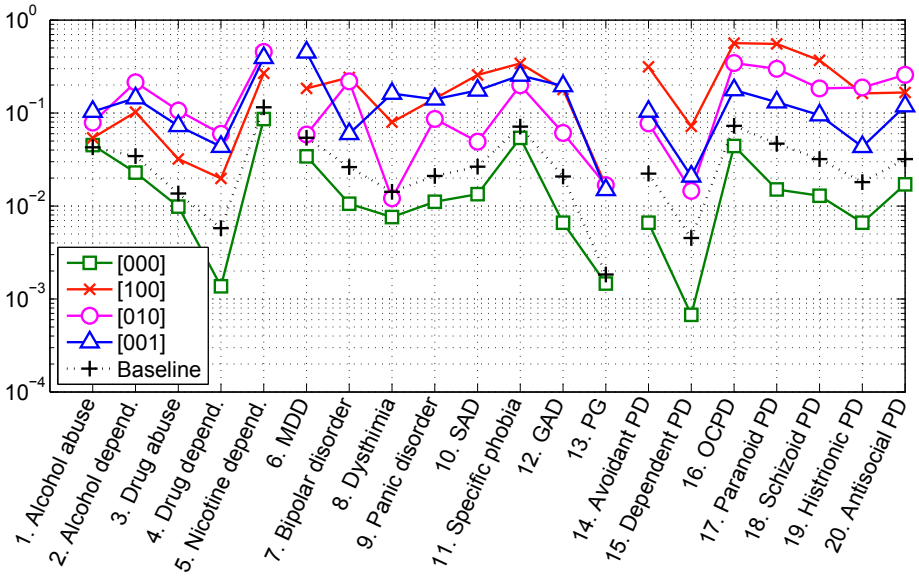


(a)

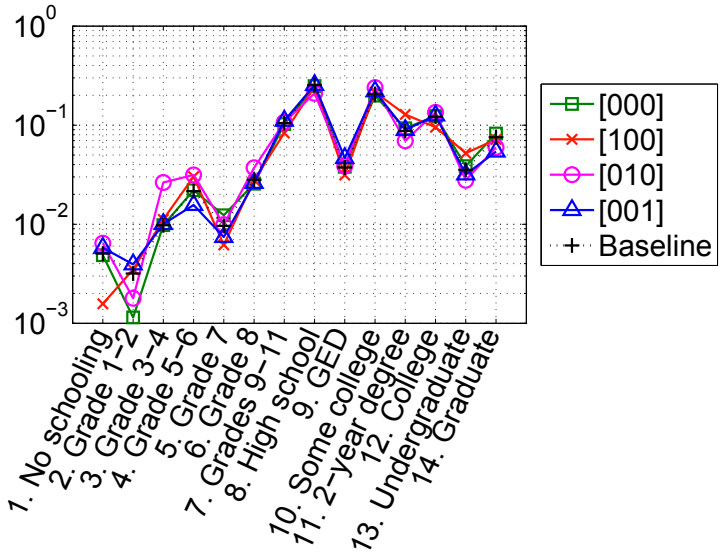


(b)

Figure 7.11: Marital Status. (a) Probabilities of suffering from the 20 considered disorders and (b) probability of belonging to each category for the latent feature vectors $\mathbf{z}_n$ shown in the legend and for the baseline.

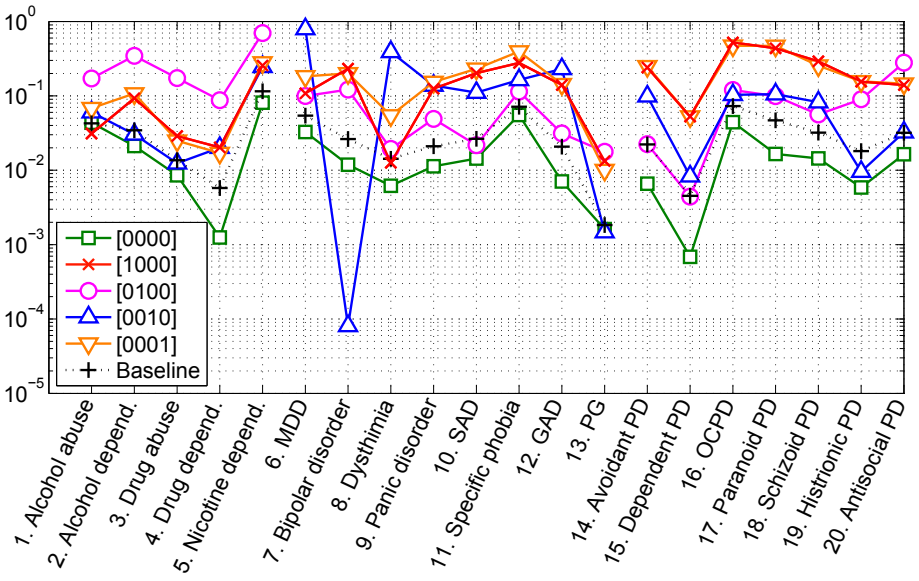


(a)

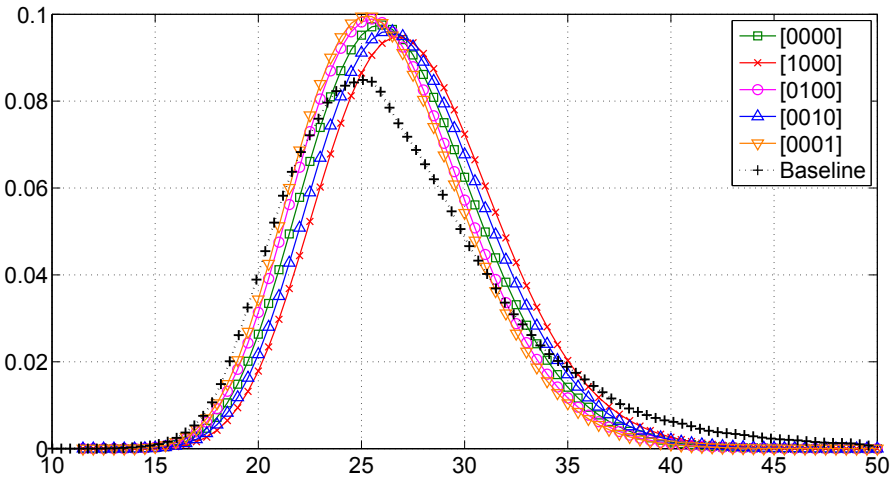


(b)

Figure 7.12: School. (a) Probabilities of suffering from the 20 considered disorders and (b) probability of belonging to each category for the latent feature vectors $\mathbf{z}_n$ shown in the legend and for the baseline.



(a)



(b)

Figure 7.13: BMI. (a) Probabilities of suffering from the 20 considered disorders for the latent feature vectors $\mathbf{z}_n$ shown in the legend; and (b) inferred probability distribution for the latent feature vectors $\mathbf{z}_n$ shown in the legend and baseline probability distribution.

Chapter 8

Analysis of Personality Disorders

8.1 Analysis of Diagnostic Criteria

Now, we study in more detail the seven personality disorders (PDs) in the previous section. In order to identify the seven personality disorders, psychiatrists have established specific diagnostic criteria for each of them. These criteria correspond to affirmative responses to one or several questions in the NESARC survey (detailed in Appendix D) and this correspondence is shown in Table 8.1. Then, there exists a set of criteria to identify if a subject presents any of the following personality disorders: avoidant, dependent, obsessive-compulsive, paranoid, schizoid, histrionic and antisocial.

In this section, we analyze how the different criteria (and their corresponding questions) are related. Our objective is to find the different comorbidity patterns in the database. With this study, we aim at answering the following three questions:

- Are the different criteria used to diagnose a disorder exchangeable (in the sense that they just indicate the PD a subject suffers from) or, on the contrary, different criteria indicate different aspects or levels of suffering from the same PD?

- Are the comorbidity patterns related to the PDs or to their criteria? We try to find out if the co-existence in a subject of two PDs is independent of the specific diagnostic criteria that the subject meets or, for instance, the probability of fulfilling a specific criterion of a PD increases when the subject meets a criterion corresponding to another

disorder.

- Are the criteria actually related to the disorders they were defined for, or some of them are more related to other PDs?

Question Code	**Personality disorder and criterion**
S10Q1A1-S10Q1B7	Avoidant (1 question per criterion)
S10Q1A8-S10Q1B15	Dependent (1 question per criterion)
S10Q1A16-S10Q1B17	OCPD criterion 1
S10Q1A18-S10Q1B23	OCPD criteria 2-7
S10Q1A24-S10Q1B25	OCPD criterion 8
S10Q1A26-S10Q1B29	Paranoid criteria 1-4
S10Q1A30-S10Q1A31	Paranoid criterion 5
S10Q1A32-S10Q1B33	Paranoid criteria 6-7
S10Q1A45-S10Q1B46	Schizoid criterion 1
S10Q1A47-S10Q1B48	Schizoid criteria 2-3
S10Q1A50-S10Q1B50	Schizoid criterion 4
S10Q1A43-S10Q1B43	Schizoid criterion 5
S10Q1A51-S10Q1B52	Schizoid criterion 6
S10Q1A49-S10Q1B49 or S10Q1A53-S10Q1B53	Schizoid criterion 7
S10Q1A54-S10Q1B54 or S10Q1A56-S10Q1B56	Histrionic criterion 1
S10Q1A58-S10Q1B58 or S10Q1A60-S10Q1B60	Histrionic criterion 2
S10Q1A55-S10Q1B55	Histrionic criterion 3
S10Q1A61-S10Q1B61	Histrionic criterion 4
S10Q1A64-S10Q1B64	Histrionic criterion 5
S10Q1A59-S10Q1B59 or S10Q1A62-S10Q1B62	Histrionic criterion 6
S10Q1A63-S10Q1B63	Histrionic criterion 7
S10Q1A57-S10Q1B57	Histrionic criterion 8
S11Q1A20-S11Q1A25	Antisocial, criterion 1
S11Q1A11- S11Q1A13	Antisocial, criterion 2
S11Q1A8- S11Q1A10	Antisocial, criterion 3
S11Q1A17- S11Q1A18	Antisocial, criterion 4
S11Q1A26- S11Q1A33	Antisocial, criterion 4
S11Q1A14- S11Q1A16	Antisocial, criterion 5
S11Q1A6 and S11Q1A19	Antisocial, criterion 6
S11Q8A-B	Antisocial, criterion 7

Table 8.1: Correspondence between the criteria for each personality disorder and questions in NESARC.

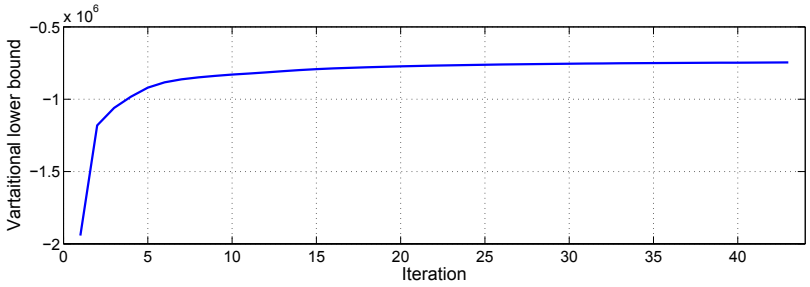


Figure 8.1: Variational lower bound $\mathcal{L}(\mathcal{H}, \mathcal{H}_q)$ at each iteration.

8.1.1 Experimental Setup

In the present analysis, we consider as input data the fulfilment of the 52 criteria (i.e., $R = 2$) corresponding to all the disorders for the 43,093 subjects and apply the variational inference algorithm truncated to $K = 25$ features, as detailed in Section 4.2, to find the latent structure of the data.

In order to properly initialize the huge amount of variational parameters, we have previously run six Gibbs samplers (combined with the Laplace approximation to compute the marginal likelihood) over the data but taking only the criteria corresponding to the avoidant PD and another PD (that is, the seven criteria for avoidant PD and the seven for dependent PD, the criteria for avoidant PD with the eight for the OCPD, etc.) for $10,000$ randomly chosen subjects. After running the six Gibbs samplers, we obtain 18 latent features that we group in a unique matrix $\mathbf{Z}$ to obtain weighting matrices $\mathbf{B}^d_{\text{MAP}}$, which are used to initialize some parameters ν_{nk} and ϕ^d_{kr}. We do this because the variational algorithm is sensitive to the starting point and a random initialization would not produce good solutions.

We run enough iterations of the variational algorithm to ensure convergence of the variational lower bound (the lower bound at each iteration is shown in Figure 8.1). We construct a binary matrix $\mathbf{Z}$ by setting each element $z_{nk} = 1$ if $\nu_{nk} > 0.5$.

8.1.2 Results

In Table 8.2, we show the probability of occurrence of each feature (top row), as well as the probability of having active only one single feature (bottom row). We also show the 'empirical' and the 'product' probabilities of possessing at least two latent features in Table 8.3, and the probabilities of possessing at least two features given that one of them is active in Table 8.4.

In Figure 8.2, we plot the probability of meeting each criterion in the general population (dashed line) and the probability of meeting each criterion for those subjects that do not have any active feature in our model (solid line). There are $15,185$ subjects (35.2% of the population) which do not present any active feature, and for these people the probability of meeting any criterion is reduced significantly.

We have found results that are in accordance with previous studies and at the same time provide new information to understand personality disorders. Out of the 10 features, 6 of them directly describe personality disorders. Feature 1 increases the probability of fulfilling the criteria for OCPD, Feature 3 increases the probability of fulfilling the criteria for antisocial, Feature 4 increases the probability of fulfilling the criteria for paranoid, Feature 5 increases the probability of meeting the criteria for schizoid, Feature 8 increases the probability of fulfilling the criteria for histrionic and Feature 7 increases the probability of meeting the criteria for avoidant and dependent.

In Figure 8.3, we plot the probability ratio between the probability of meeting each criterion when a single feature is active with respect to the probability of meeting each criterion in the general population (baseline in Figure 8.2). So, if the ratio is above one, it means that the feature increases the probability of meeting that criterion with respect to the general population. In all these plots, we also show the probability ratio between not having any active feature and the general population, which serves as a reference for a low probability of fulfilling a criterion. Note that the scale on the vertical axis may be different through all the figures for a better display. In Figure 8.3, we can see that only the criteria for one of the personality disorders is systematically above one, when one feature is active, except for Feature 7 that increases the probability for both avoidant and dependent. In the figure, we can also notice that when one feature is active the probability of the criteria for the other disorders is above the probability for the subjects that do not have any active feature, although lower than the general population (above the solid line and below one). It partially shows the comorbidity pattern for each personality disorder. For example, Feature 1, besides increasing the probability of meeting the criteria for OCPD, also increases the probability of meeting criterion 3 for schizoid and criterion 1 for histrionic. It is also important to point out that Feature 8 increases significantly the probability of meeting criteria 1, 2, 4 and 6 for histrionic (and mildly for criterion 7), but it does not affect criteria 3, 5 and 8, although the probability of meeting these criteria are increased by Feature 4 (paranoid) and Feature 5 (schizoid). In a way, it indicates that criteria 3 and 8 are more related to paranoid disorder and criterion 5 to

schizoid disorder.

As seen in Figure 8.4, Features 2 and 6 mainly reduce the probability of meeting the criteria for dependent PD. Feature 2 also reduces criteria 4-7 for avoidant and mildly increases criterion 1 for OCPD, criterion 6 for schizoid and criteria 5 and 6 for antisocial. Feature 6 also reduces some criteria below the probability for the subjects with no active features. But for most of the criteria the probability ratio moves between one and the ratio for the subjects with no active feature. When these features appear by themselves, the subjects might be similar to the subjects without any active feature, they become relevant when they appear together with other features. These features are less likely to be isolated features than the previous ones, as reported in Table 8.2. For example, Feature 2 appears frequently with Features 1, 3, 4 and 5, as shown in Table 8.4, and the probability ratios are plotted in Figure 8.5 and compared to the probability ratio when each feature is not accompanied by Feature 2. We can see that when we add Feature 2 to Feature 1, the comorbidity pattern changes significantly and it results in subjects with higher probabilities of meeting the criteria for every other disorder except avoidant and dependent. Additionally, when we add Feature 2 to Feature 5, we can see that meeting the criteria for schizoid is even more probable, together with criterion 5 for histrionic.

Either Feature 1 or Features 1 and 3 typically accompany Feature 6, and Feature 6 is seldom seen by itself (see Tables 8.2 and 8.5). In Figure 8.6, we show the probability ratio when Feature 1 is active and when Features 1 and 3 are active, as reference, and when we add Feature 6 to them. Adding Feature 6 mainly reduces the probability of meeting the criteria for dependent. It is also relevant to point out that Features 1 and 3 increase the probability of meeting the criteria 5 and 6 for paranoid, while Feature 4 mainly increased the probability of meeting the criteria 1-4 for paranoid personality disorder, as shown in Figure 8.3.

Feature 9 is similar to Feature 7, as it captures an increase in the probability of meeting the criteria for avoidant and dependent, but it never appears isolated and most times it appears together with Features 1 and 4.

Feature 10 never appears isolated and it mainly appears only with Feature 1. This feature by itself only indicates that the probability of all the criteria should be much lower than the subjects with no active features, except for antisocial, which behaves as the subjects with no active features. When we add Feature 1 to Feature 10, we get that the probability of meeting the criteria for OCDP goes to that of the subject with no active features, as can be seen in Figure 8.7. For us this is a spurious feature that is equivalent to not having any active feature and that the variational algorithm has not

been able to eliminate. This is always a risk when working with flexible models, like BNP, in which a spurious component might appear when it should not. These components can be eliminated by common sense in most cases or by further analysis by experts (psychiatric experts in our case). But it can also indicate an unknown component that can point towards a new research direction previously unknown, which is one of the attractive features of using generative models.

Besides the comorbidity patterns shown by the individual features that we have already reported, we can also see that almost all the features are positively correlated. In Table 8.3, we show the probability that any two features appear together (upper triangular sub-matrix) and the joint probability that we should observe if the features were independent (lower triangular sub-matrix). Ignoring Feature 10, all of the other features are positively correlated, except Features 2 and 7 and Features 8 and 5 that seem uncorrelated (the differences are not statistically significant). Most of the features are strongly correlated and the differences in Table 8.3 correspond to several standard deviations higher (between 3 and 42) than we should expect from independent random observations. For example, the correlation between Features 4 and 9 and Features 4 and 7 is quite high and both show subjects with higher probability of meeting the criteria for avoidant, dependent and paranoid. The difference between Features 7 and 9 is given by the criteria 1-4 for paranoid PD, that are significantly increased by Feature 9 and slightly by Feature 7, as it can be seen in Figure 8.8. Finally, it is worth mentioning that Feature 4 (paranoid) is the most highly correlated feature with all the others, so we can say that anyone suffering from paranoid PD has a higher comorbidity with any other personality disorder.

Feat.	1	2	3	4	5	6	7	8	9	10
Total	43.45	19.01	15.28	13.99	11.76	8.97	7.54	6.91	1.86	1.43
Single	13.48	3.62	2.22	1.34	2.27	0.49	0.76	1.07	0	0

Table 8.2: Probabilities (%) of possessing (top row) at least one latent feature, or (bottom row) a single feature.

Feat.	1	2	3	4	5	6	7	8	9	10
1		9.92	8.96	8.48	5.67	7.22	4.92	3.85	1.46	1.42
2	8.26		4.43	4.54	3.67	1.90	1.43	2.08	0.71	0.21
3	6.64	2.90		3.29	2.18	3.00	2.02	1.58	0.54	0.20
4	6.08	2.66	2.14		2.79	1.91	2.39	1.40	1.25	0.03
5	5.11	2.23	1.80	1.65		1.31	1.35	0.85	0.57	0.00
6	3.90	1.71	1.37	1.26	1.05		1.10	0.80	0.44	0.14
7	3.28	1.43	1.15	1.06	0.89	0.68		0.65	0.28	0.00
8	3.00	1.31	1.06	0.97	0.81	0.62	0.52		0.51	0.07
9	0.81	0.35	0.28	0.26	0.22	0.17	0.14	0.13		0.00
10	0.62	0.27	0.22	0.20	0.17	0.13	0.11	0.10	0.03	

Table 8.3: Probabilities (%) of possessing at least two latent features. The elements above the diagonal correspond to the 'empirical probability', that is, extracted directly from the inferred IBP matrix $\mathbf{Z}$, and the elements below the diagonal correspond to the 'product probability' of the corresponding two latent feature probabilities given in the first row of Table 8.2.

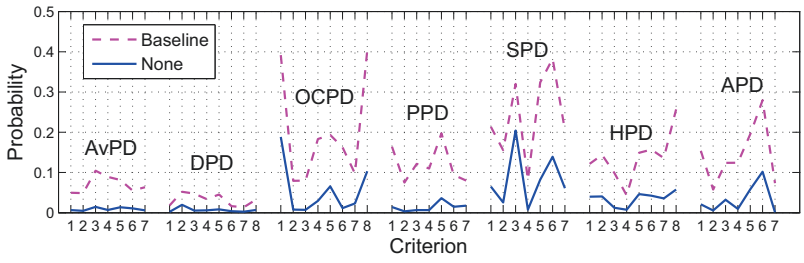


Figure 8.2: Probability of meeting each criterion. The probabilities when no latent feature is active (solid curve) have been obtained using the matrices $\mathbf{B}^d_{\text{MAP}}$, while the baseline (dashed curve) has been obtained taking into account the $43,093$ subjects in the database. (AvPD=Avoidant PD, DPD=Dependent PD, OCPD=Obsessive-compulsive PD, PPD=Paranoid PD, SPD=Schizoid PD, HPD=Histrionic PD, APD=Antisocial PD)

k_1 \ k_2	1	2	3	4	5	6	7	8	9	10
1	100	22.83	20.63	19.53	13.05	16.62	11.33	8.85	3.37	3.27
2	52.19	100	23.33	23.90	19.32	10.00	7.51	10.95	3.75	1.09
3	58.68	29.03	100	21.54	14.29	19.66	13.25	10.34	3.51	1.29
4	60.63	32.47	23.52	100	19.97	13.65	17.05	10.02	8.92	0.20
5	48.22	31.25	18.57	23.77	100	11.11	11.49	7.24	4.88	0.00
6	80.47	21.18	33.47	21.29	14.56	100	12.23	8.92	4.86	1.53
7	65.26	18.92	26.83	31.63	17.91	14.55	100	8.65	3.66	0.03
8	55.62	30.11	22.86	20.28	12.32	11.58	9.43	100	7.39	1.07
9	78.46	38.23	28.77	67.00	30.76	23.41	14.82	27.40	100	0.12
10	99.19	14.40	13.75	1.94	0.00	9.55	0.16	5.18	0.16	100

Table 8.4: Probabilities (%) of possessing at least features k_1 and k_2 given that k_1 is active, i.e., $\left(\sum_{n=1}^{N} z_{nk_1} z_{nk_2}\right) / \left(\sum_{n=1}^{N} z_{nk_1}\right)$.

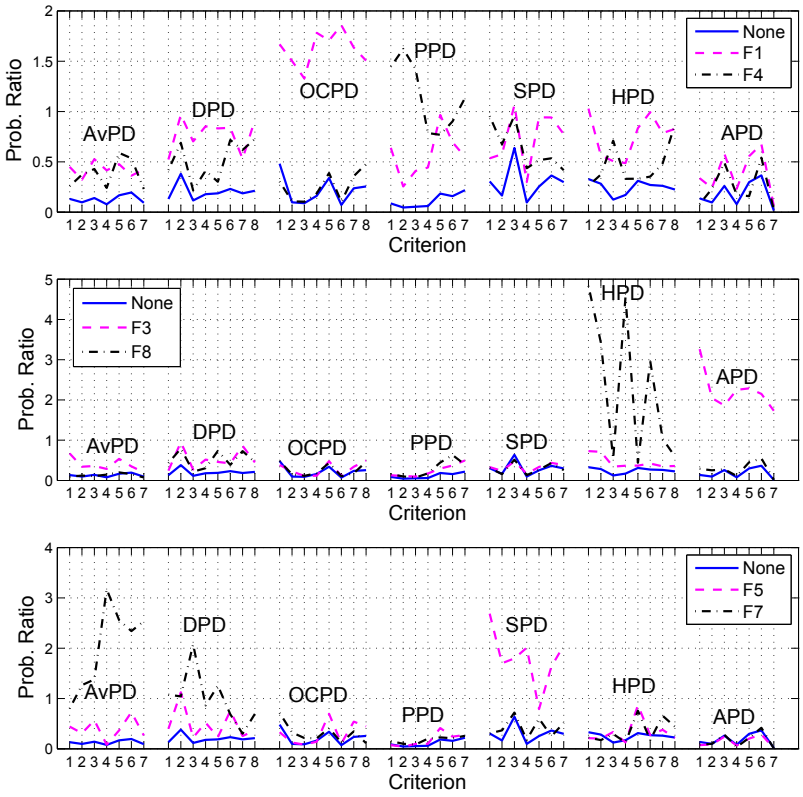


Figure 8.3: Probability ratio of meeting each criterion, with respect to the baseline. These probabilities have been obtained using the matrices $\mathbf{B}^d_{\text{MAP}}$, when none or a single feature is active (the legend shows the active latent features).

	# Occurrences	Features									
		1	2	3	4	5	6	7	8	9	10
1	15185	0	0	0	0	0	0	0	0	0	0
2	5811	1	0	0	0	0	0	0	0	0	0
3	1561	0	1	0	0	0	0	0	0	0	0
4	1389	1	1	0	0	0	0	0	0	0	0
5	1021	1	0	1	0	0	0	0	0	0	0
6	977	0	0	0	0	1	0	0	0	0	0
7	958	1	0	0	0	0	1	0	0	0	0
8	956	0	0	1	0	0	0	0	0	0	0
9	946	1	0	0	1	0	0	0	0	0	0
10	687	1	0	0	0	1	0	0	0	0	0
11	576	0	0	0	1	0	0	0	0	0	0
12	553	1	0	0	0	0	0	1	0	0	0
13	495	0	1	0	0	1	0	0	0	0	0
14	486	1	0	0	0	0	0	0	1	0	0
15	460	0	0	0	0	0	0	0	1	0	0
16	451	0	1	1	0	0	0	0	0	0	0
17	438	1	0	0	0	0	0	0	0	0	1
18	414	1	0	1	0	0	1	0	0	0	0
19	385	0	1	0	1	0	0	0	0	0	0
20	370	1	1	0	1	0	0	0	0	0	0

Table 8.5: List of the 20 most common feature patterns.

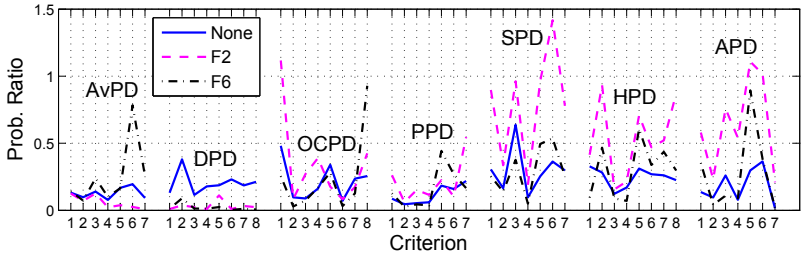


Figure 8.4: Probability ratio of meeting each criterion, with respect to the baseline. These probabilities have been obtained using the matrices $\mathbf{B}^d_{\text{MAP}}$, when none or a single feature is active (the legend shows the active latent features).

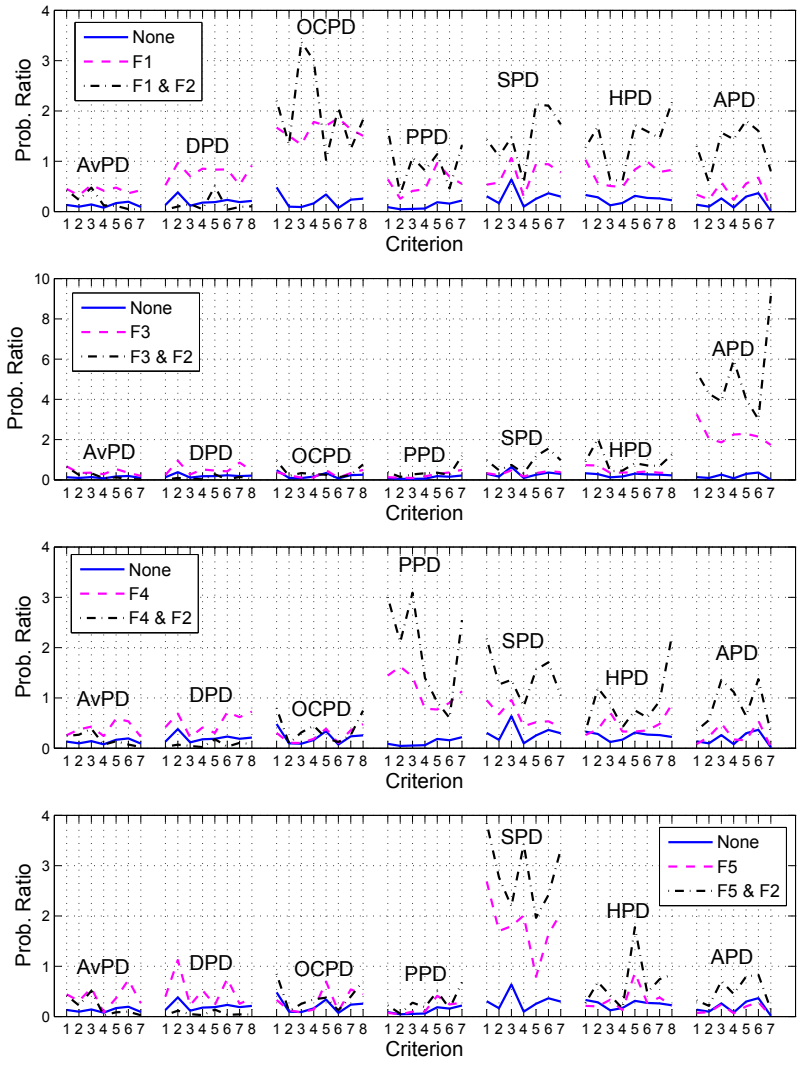


Figure 8.5: Probability ratio of meeting each criterion, with respect to the baseline. These probabilities have been obtained using the matrices $\mathbf{B}^d_{\text{MAP}}$, when none, a single or two features are active (the legend shows the active latent features).

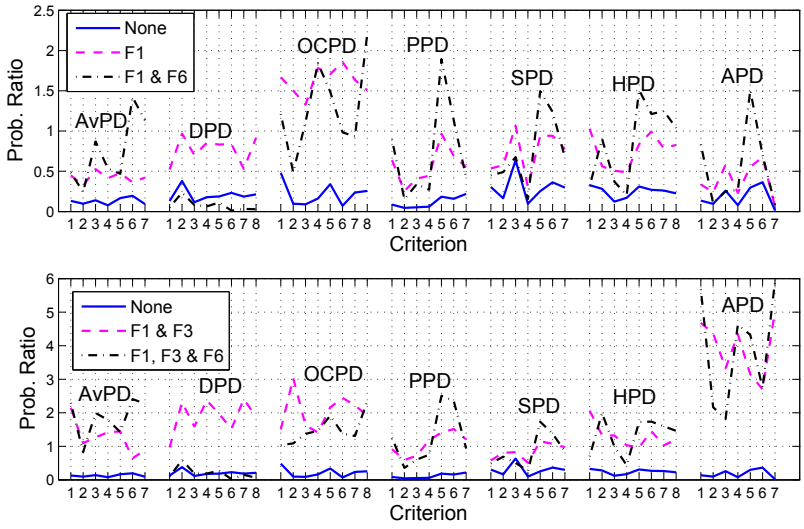


Figure 8.6: Probability ratio of meeting each criterion, with respect to the baseline. These probabilities have been obtained using the matrices $\mathbf{B}^d_{\text{MAP}}$, when none, a single or several features are active (the legend shows the active latent features).

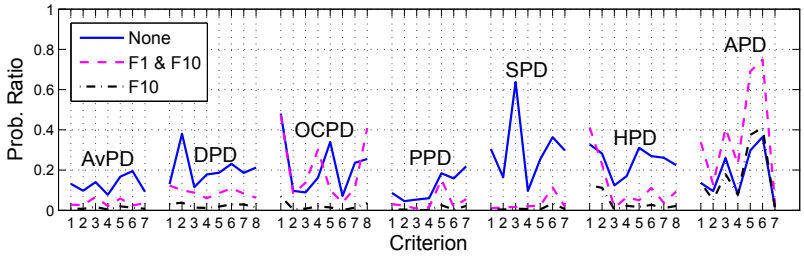


Figure 8.7: Probability ratio of meeting each criterion, with respect to the baseline. These probabilities have been obtained using the matrices $\mathbf{B}^d_{\text{MAP}}$, when none, a single or two features are active (the legend shows the active latent features).

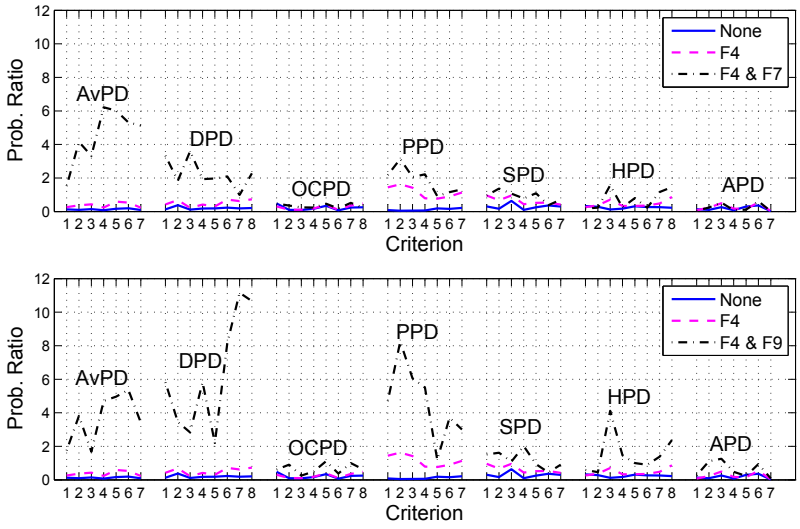


Figure 8.8: Probability ratio of meeting each criterion, with respect to the baseline. These probabilities have been obtained using the matrices $\mathbf{B}^d_{\text{MAP}}$, when none, a single or two features are active (the legend shows the active latent features).

8.2 Analysis of the Survey Responses

In the previous study, we have worked with the results obtained after processing the responses to the survey, i.e., the fulfilment of the criteria. As detailed above, these criteria correspond to affirmatively answering to one or a set of questions (see Table 8.1). According to the codebook of the NESARC database, the questions are organized in pairs: there is a first question (coded in Table 8.1 as S10Q1AX) with three possible responses, 'yes', 'no' and 'unknown'; and a second question (coded in Table 8.1 as S10Q1BX) that is asked only in case the subject has responded affirmatively to the former. Hence, for each pair of questions, we find five possible outcomes: 'no', 'unknown', 'yes+no', 'yes+unknown' and 'yes+yes'. Based on these responses, the psychiatrists consider that a subject meets a criterion if the subject has answered affirmatively to the first question of the pair, i.e., if she has responded 'yes+no', 'yes+unknown' or 'yes+yes' to the pair of questions that define a criterion. Moreover, for those criteria with more than one associated (pairs of) questions, they assume that the subject satisfies the criterion if she has answered 'yes+no', 'yes+unknown' or 'yes+yes' to any

(of the pairs) of questions.

However, based on the previous results, we believe that there is further information about the comorbidity patterns among psychiatric disorders in the responses to the survey. As a consequence, in this section we work directly with the responses of the people to the survey with the aim of answering the following set of questions:

- Are the different questions designed to diagnose each PD actually related to the disorder they were defined for, or some of them are more related to another PD?
- Do the subjects that suffer from different disorders respond in a different way to the questions? In other words, do the different disorders present different response patterns?
- How are the questions related among them?

8.2.1 Experimental Setup

In this study, we use the responses to Section 10 in NESARC database as input data to the IBP. This section of the NESARC contains 55 pairs of questions used to diagnose the six following PDs: avoidant, dependent, obsessive-compulsive, paranoid, schizoid and histrionic. We consider each pair of questions as a unique input which takes five possible values: 'no', 'unknown', 'yes+no', 'yes+unknown' and 'yes+yes'.

For this experiment, we resort to the model and inference detailed in Chapter 5, assuming that all the attributes in the database are categorical with five categories. We set $\alpha = 1$, $\sigma_y^2 = 1$ and $\sigma_B^2 = 1$, and run the inference algorithm detailed in Section 5.2. In order to get more interpretable results, we do not sample the rows of $\mathbf{Z}$ corresponding to those subjects who responded negatively to the 55 questions but instead fix these latent features to zero. The idea is that the bias terms capture the general population, and we use the active components of the matrix $\mathbf{Z}$ to characterize the disorders.

8.2.2 Results

After running our inference algorithm, we obtain eight latent variables. In Table 8.6, we show the probability of occurrence of each feature (top row), as well as the probability of having active only one single feature (bottom row). In this table, we observe that there are two groups of latent features: the first two most common features that are active in more than 35% of the

people and appear half of the times as unique features; and the remaining six features that are active only in a few subjects and rarely appear as unique features.

In Figure 8.9, we plot the probability of answering 'no' to each question in the general population (dashed line) and the probability of answering 'no' to each question for those subjects that do not have any active feature in our model (solid line). We obtain that 34.81% of the population do not have any active feature and, therefore, their answers are explained with the bias term. This result is in agreement with the ones obtained in the previous study, where 35.2% of the population was also explained with the bias term. As expected, people with no active latent features present higher probability of answering negatively to all the questions than the general population represented with the dashed line, being these probabilities higher than 0.9 for all the questions.

Now, we focus on Features 1 and 2, which are the most active features in the population. In Figure 8.10, we show the probability ratio between the probability of the response 'yes+no' for all the questions when Feature 1 appears as the unique active feature with respect to the probability of the 'yes+no' response in the general population. Similarly, we plot in Figure 8.11 the probability ratio between the probability of the responses 'yes+no' and 'unknown' when only Feature 2 is active with respect to the probability of the 'yes+no' and 'unkown' responses in the general population. Note that, if the ratio is above one, it means that the feature increases the probability of this response with respect to the general population. The different scales on the vertical axis provide a better display. In these figures, we observe that subjects that have active either Feature 1 or Feature 2 correspond to people that do not suffer from any disorders but have responded 'yes+no' to some pairs of questions mainly related to, respectively, schizoid PD (questions 1, 5 and 9) and obsessive compulsive PD (questions 1 and 10). Additionally, Feature 2 increases the probability of answering 'unknown' to question 10 of obsessive compulsive PD and question 4 of histrionic DP.

In Figure 8.12, we show the probability ratio between the probability of the responses 'yes+no' and 'unknown' for all the questions when Feature 3 appears as the unique active feature with respect to the probability of the 'yes+no' and 'unknown' responses in the general population. We find in Figure 8.12 that subjects with Feature 3 active present higher probability than the general population of answering 'yes+no' to some questions (1, 4 and 5) related to histrionic PD; and 'unknown' to questions 9 and 10 of OCPD, and question 4 of HPD.

In Figure 8.13, we show the probability ratio for the response 'unknown'

when Feature 4 appears as the unique active feature. Clearly, Feature 4 models those subjects that answer 'unknown' to all the questions, i.e., those that did not want to respond to the survey.

In Figure 8.14, we show the probability ratio for the response 'yes+yes' when Feature 5 appears as the unique active feature. Feature 5 captures an increase up to 10 times in the probability of answering affirmatively to the questions related to avoidant PD (and dependent PD). In addition, we observe in this figure that people that suffer from avoidant PD and dependant PD also tend to answer affirmatively to question 9 of paranoid PD, question 2 of schizoid PD, and question 10 of histrionic PD.

In Figure 8.15, we show the probability ratio for the responses 'yes+yes' and 'unknown' for all the questions when Feature 6 appears as the unique active feature. Feature 6 mainly captures the affirmative (and also 'unknown', although to a lesser extent) to the questions designed to diagnose paranoid PD. It also increases the probability of answering 'yes+yes' (or 'unknown') to some of the questions related to obsessive compulsive (questions 5, 7, 9 and 10) and histrionic PDs (questions 2 and 4).

In Figure 8.16, we show the probability ratio between the probability of the response 'yes+yes' when Feature 7 appears as the unique active feature with respect to the probability of answering 'yes+yes' in the general population. Feature 7 captures an increase in the probability of the affirmative response (i.e., 'yes+yes') for all the questions, and specially, in the questions related to avoidant, dependent and paranoid PDs. Therefore, a subject with Feature 7 active suffers from several PDs and, according to the results in Chapter 7, would present a high grade or severity of suffering of personality disorders.

In Figure 8.17, we show the probability ratio between the probability of the response 'yes+no' when Feature 8 appears as the unique active feature with respect to the probability of answering 'yes+no' in the general population. Feature 8 captures an increase in the probability of the 'yes+no' response for the questions related to avoidant, paranoid and schizoid PDs. This feature models those subjects that respond affirmatively to the first question of the pair of questions and negatively to the second question, and therefore, those subject with a moderate suffering of several disorders.

Additionally, in Table 8.7, we show the 20 most common feature patterns in the database, which capture 98.65% of the population. We have divided this table into two groups of features: the first group with Features from 1 to 4, which are the most common features and model responses 'yes+no' and 'unknown'; and a second group with Features from 5 to 8, which model

the PDs. In this table, we observe that Feature 8 does not appear in any of the most common patterns. There are 12 patterns (patterns 1-9, 12, 14 and 17) for which only features in the first group are active. In the remaining 8 feature patterns, one of the latent features in the second group (i.e., Features 5, 6 and 7) are combined with the features in the first group to model the subjects with higher risk of suffering from one or several personality disorders.

As a summary, we find that besides the 34.81% of the population without any active feature, there is another 38% of the population (see bottom row in Table 8.6), corresponding to the subjects that have Features 1, 2 or 3 as unique feature, that do not suffer from any disorder but have higher probability of answering 'yes+no' to some of the questions. Feature 5 models the subjects that suffer from avoidant and dependent PDs, and Feature 6 models paranoid PD with obsessive compulsive and histrionic tendencies. In contrast to the previous section, we only obtain a latent feature to model the affirmative responses (i.e., 'yes+yes') for all the questions associated to avoidant and dependant PDs, and paranoid PD. For the remaining disorders (i.e., obsessive compulsive, schizoid, and histrionic PDs), we find that the 'yes+yes' and 'yes+no' responses are modeled in general by the same latent variable. Finally, we remark that, as shown by all the results in this chapter, the comorbidity patterns are more related to different aspects (criteria or questions) that characterize the disorders than to the PDs themselves.

Feat.	1	2	3	4	5	6	7	8
Total	38.53	36.78	7.83	3.33	2.96	1.15	0.67	0.45
Single	19.09	17.82	1.35	1.12	0.002	0.12	0.06	0.04

Table 8.6: Probabilities (%) of possessing (top row) at least one latent feature, or (bottom row) a single feature.

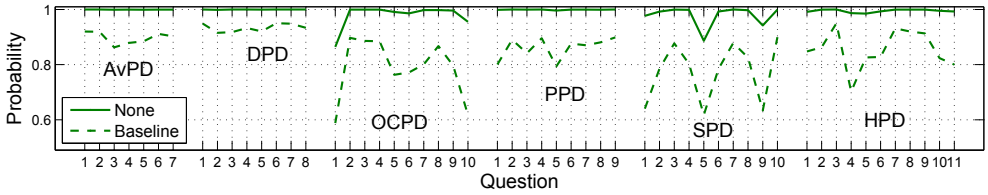


Figure 8.9: Probability of answering 'NO' to each question. The probabilities when no latent feature is active (solid curve) have been obtained using the inferred matrices $\mathbf{B}^d$, while the baseline (dashed curve) has been obtained taking into account the $43,093$ subjects in the database. AvPD=Avoidant PD, DPD=Dependent PD, OCPD=Obsessive-compulsive PD, PPD=Paranoid PD, SPD=Schizoid PD, HPD=Histrionic PD.

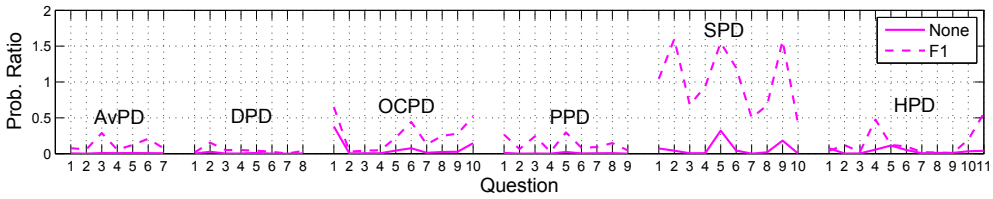
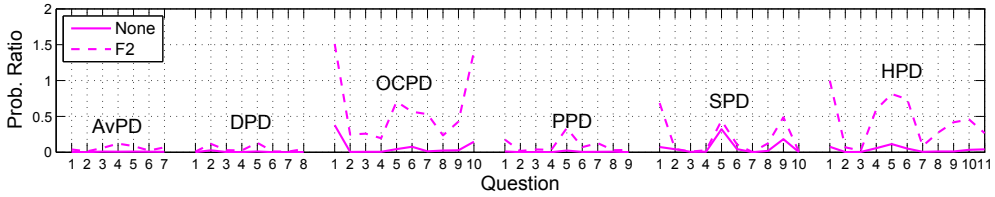
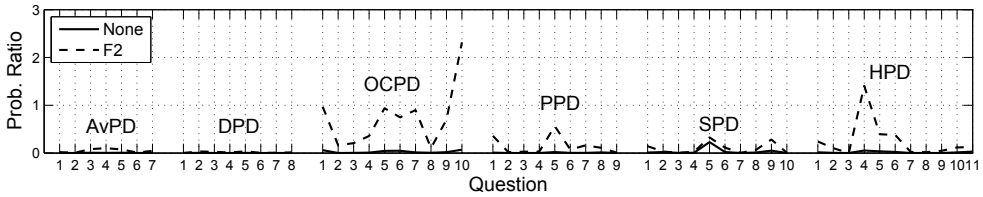


Figure 8.10: Probability ration of answering 'YES+NO' to each question with respect to the baseline. The probabilities when none or only one latent feature is active have been obtained using the inferred matrices $\mathbf{B}^d$, while the baseline has been obtained taking into account the $43,093$ subjects in the database. AvPD=Avoidant PD, DPD=Dependent PD, OCPD=Obsessive-compulsive PD, PPD=Paranoid PD, SPD=Schizoid PD, HPD=Histrionic PD.

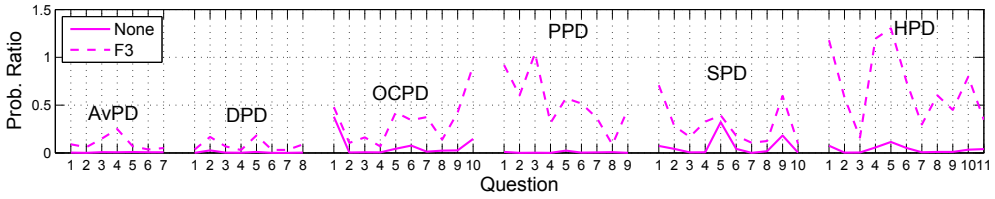


(a) 'YES+NO'

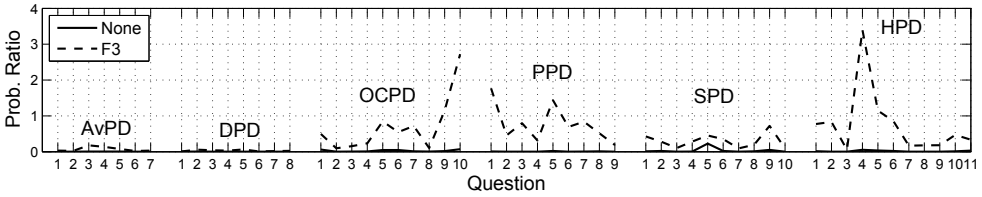


(b) 'UNKNOWN'

Figure 8.11: Probability ratio of answering 'YES+NO' and 'UNKNOWN' to each question, with respect to the baseline. The probabilities when none or only one latent feature is active have been obtained using the inferred matrices $\mathbf{B}^d$, while the baseline has been obtained taking into account the $43,093$ subjects in the database. AvPD=Avoidant PD, DPD=Dependent PD, OCPD=Obsessive-compulsive PD, PPD=Paranoid PD, SPD=Schizoid PD, HPD=Histrionic PD.



(a) 'YES+NO'



(b) 'UNKNOWN'

Figure 8.12: Probability ratio of answering 'YES+NO' and 'UNKNOWN' to each question, with respect to the baseline. The probabilities when none or only one latent feature is active have been obtained using the inferred matrices $\mathbf{B}^d$, while the baseline has been obtained taking into account the $43,093$ subjects in the database. AvPD=Avoidant PD, DPD=Dependent PD, OCPD=Obsessive-compulsive PD, PPD=Paranoid PD, SPD=Schizoid PD, HPD=Histrionic PD.

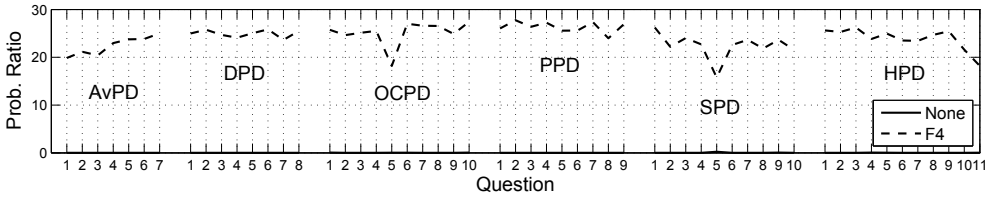


Figure 8.13: Probability ratio of answering 'UNKNOWN' to each question, with respect to the baseline. The probabilities when none or only one latent feature is active have been obtained using the inferred matrices $\mathbf{B}^d$, while the baseline has been obtained taking into account the $43,093$ subjects in the database. AvPD=Avoidant PD, DPD=Dependent PD, OCPD=Obsessive-compulsive PD, PPD=Paranoid PD, SPD=Schizoid PD, HPD=Histrionic PD.

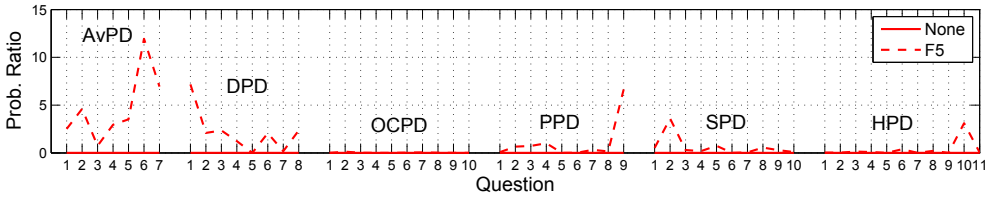
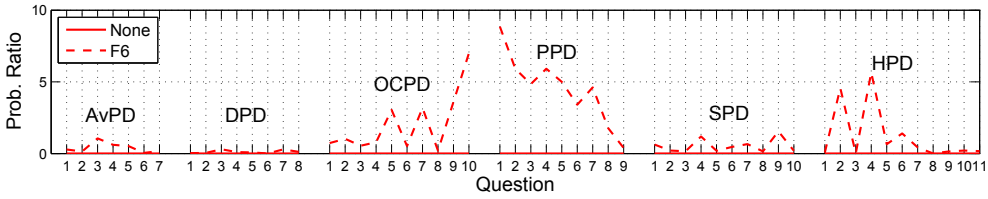
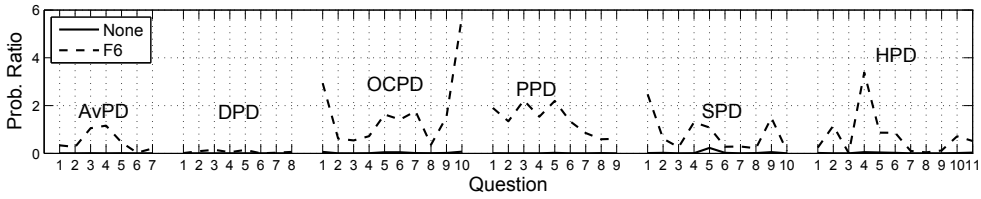


Figure 8.14: Probability ratio of answering 'YES+YES' to each question, with respect to the baseline. The probabilities when none or only one latent feature is active have been obtained using the inferred matrices $\mathbf{B}^d$, while the baseline has been obtained taking into account the $43,093$ subjects in the database. AvPD=Avoidant PD, DPD=Dependent PD, OCPD=Obsessive-compulsive PD, PPD=Paranoid PD, SPD=Schizoid PD, HPD=Histrionic PD.



(a) 'YES+YES'



(b) 'UNKNOWN'

Figure 8.15: Probability ratio of answering 'YES+YES' and 'UNKNOWN' to each question, with respect to the baseline. The probabilities when none or only one latent feature is active have been obtained using the inferred matrices $\mathbf{B}^d$, while the baseline has been obtained taking into account the $43,093$ subjects in the database. AvPD=Avoidant PD, DPD=Dependent PD, OCPD=Obsessive-compulsive PD, PPD=Paranoid PD, SPD=Schizoid PD, HPD=Histrionic PD.

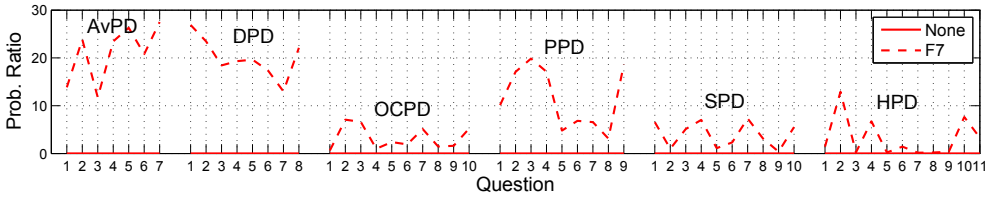


Figure 8.16: Probability ratio of answering 'YES+YES' to each question, with respect to the baseline. The probabilities when none or only one latent feature is active have been obtained using the inferred matrices $\mathbf{B}^d$, while the baseline has been obtained taking into account the $43,093$ subjects in the database. AvPD=Avoidant PD, DPD=Dependent PD, OCPD=Obsessive-compulsive PD, PPD=Paranoid PD, SPD=Schizoid PD, HPD=Histrionic PD.

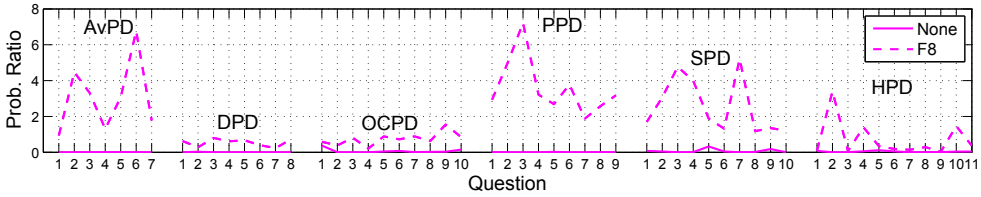


Figure 8.17: Probability ratio of answering 'YES+NO' to each question, with respect to the baseline. The probabilities when none or only one latent feature is active have been obtained using the inferred matrices $\mathbf{B}^d$, while the baseline has been obtained taking into account the $43,093$ subjects in the database. AvPD=Avoidant PD, DPD=Dependent PD, OCPD=Obsessive-compulsive PD, PPD=Paranoid PD, SPD=Schizoid PD, HPD=Histrionic PD.

	# Occurrences	Features							
		1	2	3	4	5	6	7	8
1	15002	0	0	0	0	0	0	0	0
2	8225	1	0	0	0	0	0	0	0
3	7679	0	1	0	0	0	0	0	0
4	5394	1	1	0	0	0	0	0	0
5	1224	1	0	1	0	0	0	0	0
6	1162	0	1	1	0	0	0	0	0
7	581	0	0	1	0	0	0	0	0
8	499	1	0	0	1	0	0	0	0
9	478	0	0	0	1	0	0	0	0
10	457	0	1	0	0	1	0	0	0
11	376	1	0	0	0	1	0	0	0
12	268	0	1	0	1	0	0	0	0
13	207	0	0	0	0	1	0	0	0
14	177	1	1	1	0	0	0	0	0
15	169	1	0	0	0	0	1	0	0
16	167	0	1	0	0	0	1	0	0
17	162	1	1	0	1	0	0	0	0
18	116	0	1	0	0	0	0	1	0
19	84	1	1	0	0	1	0	0	0
20	84	1	0	0	0	0	0	1	0

Table 8.7: List of the 20 most common feature patterns.

Chapter 9

Summary and Conclusions

9.1 Summary and Final Remarks

In this section, we summarize the main ideas and findings exposed throughout the thesis. We also provide some final remarks that open the future lines of research detailed in Section 9.2. Similarly to the Introduction, we split the summary into two parts: a first part that contains a summary of the Bayesian nonparametric algorithms proposed in the thesis, and a second part with a summary of the results.

9.1.1 Technical Details

In this section, we provide a summary of Chapters 3 to 5 of the thesis, i.e., a summary of the technical contributions of this thesis.

IBP model for Categorical Observations

In Chapter 3, we have extended the IBP model to deal with categorical observations. Specifically, we have considered two likelihood observation models (a multinomial-logit and a multinomial-probit model) and, motivated by our specific application, we have extended the IBP prior in two ways: i) We have included a bias term; and ii) we have extended the model to account for bounded real-valued latent variables, instead of on-off latent features. For the proposed models, we have derived in Chapter 4 several MCMC based inference algorithms and a variational inference algorithm. Specifically, we have proposed an (approximated) collapsed Gibbs sampler in which the marginal likelihood is approximated using either the Laplace

approximation or the EP algorithm under, respectively, the multinomial-logit or the multinomial-probit model. Additionally, we have proposed an MH based algorithm to infer the latent variables in the continuous feature model.

Note that, although our work has been motivated by our specific psychiatric application, the proposed models and the corresponding inference algorithms are general enough to be applicable in any context dealing with categorical observations. We also remark that, although both approximations present linear complexity with the number of observations, the EP algorithm presents higher computational cost than the Laplace approximation because it needs several iterations of both the inner and outer loops at each step of the Gibbs sampler algorithm. But, in turn, the EP algorithm also provides more accurate estimates of the marginal likelihood.

IBP model for Heterogeneous Observations

In Chapter 5, we have proposed a general observation model for the IBP that allows us to handle mixed continuous and discrete variables. More specifically, the proposed model is able to manage real-valued, positive real-valued, categorical, ordinal, and count data. For this model, we have derived an MCMC inference algorithm, based on the accelerated Gibbs sampler for the IBP [18], that scales linearly with the number of observations. This algorithm performs exact inference, being the computation of marginal likelihood analytically tractable, by introducing an auxiliary Gaussian variable such that, conditioned on this variable, it resembles the standard Gaussian IBP model.

This model provides an efficient and general Bayesian approach for applying probabilistic modeling to heterogeneous databases, which are very common in real applications. Finally, note that the proposed model when dealing with categorical observations coincides with the one in Chapter 3 under the multinomial-probit likelihood function but, in contrast to the inference algorithm proposed in Chapter 4, the introduction of the auxiliary Gaussian variable allows for exact inference. However, although both algorithms (the ones in Chapter 4 and the one in Chapter 5), collapse the weighting factors by computing (either approximately or exactly) the marginal likelihood, the introduction of an auxiliary variable (that needs to be sampled) may deteriorate the mixing performance of the algorithm. Therefore, an extensive study of the mixing properties of the proposed algorithms (i.e., the approximate and exact collapsed Gibbs samplers and the variational inference algorithm) appears as an interesting future research

line.

9.1.2 Experiments

In this section, we provide a summary of the main results obtained in the second part of the thesis, i.e., in Chapters 6 to 8.

Analysis of Suicide Attempts

In this study, we have applied the IBP model for categorical observations (under the multinomial-logit likelihood observation model) to the NESARC database to find the hidden features that characterize the suicide attempt risk. From the analysis of how each inferred feature contributes to the suicide attempt probability, we have found that our algorithm is able to detect the people with the highest and the lowest risk of attempting suicide.

Let us remark that the proposed approach can be used to discard significant portions of the population in suicide attempt studies and focus on the groups that present much higher risk. Hence, our IBP for categorical observations is able to obtain features that describe the hidden causes behind suicide attempts and makes it possible to pin-point the people that have a higher risk of attempting suicide.

Analysis of Psychiatric disorders

In Chapter 7, we have used the diagnoses of the 20 psychiatric disorders available in the NESARC database to perform a thorough analysis of the comorbidity patterns among these disorders. In this study, we have considered the continuous latent feature model, in which the latent variables take bounded real values. We have shown that the obtained results are not only consistent with previous studies on the latent structure of psychiatric disorders but also provide new insights. We have found that the comorbidity patterns of common psychiatric disorders can be described by a small number of latent features, even though the model has enough a priori flexibility to account for a potentially unbounded number of features. In addition, nosologically related disorders, such as social anxiety disorder and avoidant personality disorder, tend to be modeled by similar features. We have found that no disorder is perfectly aligned along one single latent feature, which suggests that disorders can develop through multiple etiological paths. For instance, the risk of nicotine dependence may be high in individuals with a propensity towards externalization or internalization, as suggested in [7]. We have observed that the 20 psychiatric

disorders under study can be divided into three groups of latent disorders, namely internalizing, externalizing and personality disorders. Furthermore, comorbid disorders tend to be modeled by the same latent feature, i.e., tend to belong to the same group of latent disorders. The importance of the severity factors in the model has also been proved, because they allow explaining the comorbidity among the disorders and also understanding the stress each subject suffers. The model without the severity factors cannot distinguish between the different subjects that have the same active latent features.

Then, we have made use of the IBP model for heterogeneous databases to study the impact of the social background of the subjects in their comorbidity patterns. In particular, we have studied how sex, age, census region, race/ethnicity, marital status, highest grade or years of school completed, and the BMI show up in the comorbidity patterns among the 20 considered psychiatric disorders. In this study, we have found that, in agreement with previous studies, women tend to suffer from mood and anxiety disorders (internalizing factor) in a higher extent than men, who frequently suffer from personality disorders. Additionally, we have found that the body max index (BMI) also influence the development of some latent disorders, finding that people with larger BMI tend to suffer in a higher extent from mood and anxiety disorders.

Analysis of personality disorders

In Chapter 8, we have performed a thorough analysis of the comorbidity patterns among the seven PDs diagnosed using the data in the NESARC. For this analysis, we have worked with the criteria that the psychiatrists defined to diagnose these seven PDs, instead the diagnoses themselves. We have found a latent feature to directly describe each personality disorder, except the avoidant and dependent PDs that are modeled by the same latent feature. We also found that paranoid PD is the most highly correlated PD with all the others, so we can say that anyone suffering from paranoid PD has a higher probability of suffering from comorbid PDs.

Afterwards, we have studied directly the responses to the NESARC survey, instead of the fulfilment of the diagnostic criteria obtained after processing the data. In this analysis, we have observed that approximately 38% of the population answer 'yes+no' to some of the pairs of questions used to diagnose PDs. This makes us wonder if the way these questions are stated provides useful information to detect those subjects that suffer from PDs, or they should be reformulated in a way that provide the information we are

looking for. Additionally, we have observed that the comorbidity patterns are more related to specific questions rather than to PDs. For instance, once a subject suffers from avoidant and dependent PDs, she also affirmatively answers to some specific questions of paranoid, schizoid and histrionic PDs.

Another question that arises from our results is related to the avoidant and dependent PDs, which in all the performed experiments are modeled by a unique latent feature, which, in agreement with previous studies [24], indicates that they are highly correlated. Therefore, we may wonder whether avoidant and dependent PDs are two different PDs or, on the contrary, they correspond to different levels of suffering from the same PD.

9.2 Future Work

Several extensions of this work can be performed in both machine learning and psychiatry. On the one hand, we have possible future research lines regarding the Bayesian nonparametric models and inference algorithms. As we have pointed out before, the proposed models and inference algorithms are general enough to be applicable in other areas distinct from psychiatry. For instance, possible extensions of this thesis include: i) The development of an alternative and more scalable inference algorithm for the proposed continuous latent feature model; ii) a throughout analysis of the mixing properties of the proposed MCMC based inference algorithms; or iii) the derivation of a variational inference algorithm that, instead of bounding the lower bound, directly approximates the lower bound, which would probably provide better results [90]. Additionally, an interesting extension of the IBP model for heterogeneous databases is the development of a general tool for the estimation of missing data in such databases. Specifically, this model is able to directly provide estimates of the missing data by exploiting the information in the available data to learn the similarities among the objects and how these latent features show up in the attributes that describe the objects. Finally, the idea of introducing an auxiliary Gaussian variable, i.e., a pseudo-observation, could be combined with other Bayesian models, e.g., with the DP mixture model to perform clustering in heterogeneous databases.

On the other hand, the exhaustive analysis of different problems in the area of psychiatry has led to a set of open questions. Hence, the search of the responses to these questions appears as a natural future line of research in this area. In this work, we have focused on the study of the causes behind suicide attempts and the comorbidity patterns of psychiatric disorders, but

the NESARC database contains further information that can be used to study other problems. For instance, we could study the hidden causes behind substance use and abuse disorders. The NESARC database would also allow for socioeconomical studies because, in addition to the information of the mental health of the participants, it also contains information of the social background of the participants such as their incomes or ethnicity, among others. Finally, in order to study how the different disorders evolve with time, we would need temporal information of the subjects. This study could be analyzed by psychiatrists to better understand the different phases of a disorder and, as a consequence, help them to detect and treat the subjects beforehand, avoiding visits to the emergency rooms or even preventing suicide attempts.

Appendix A

Laplace Approximation

In this chapter we provide the necessary details for the implementation of the Laplace approximation proposed in Section 4.1.1. The expression in (4.6) can be rewritten as

$$f(\mathbf{B}^d) = \text{trace}\left\{\mathbf{M}^{d\top}\mathbf{B}^d\right\} - \sum_{n=1}^{N}\log\left(\sum_{r=1}^{R}\exp(\mathbf{z}_n\mathbf{b}_{\cdot r}^d)\right) - \frac{1}{2\sigma_B^2}\text{trace}\left\{\mathbf{B}^{d\top}\mathbf{B}^d\right\} - \frac{RK}{2}\log(2\pi\sigma_B^2),$$

where $(\mathbf{M}^d)_{kr}$ counts the number of data points for which $x_{nd} = r$ and $z_{nk} = 1$, namely, $(\mathbf{M}^d)_{kr} = \sum_{n=1}^{N}\delta(x_{nd} = r)z_{nk}$, where $\delta(\cdot)$ is the Kronecker delta function. By definition, $(\mathbf{M}^d)_{0r} = \sum_{n=1}^{N}\delta(x_{nd} = r)$.

By defining $(\boldsymbol{\rho}^d)_{kr} = \sum_{n=1}^{N} z_{nk}\pi_{nd}^r$, the gradient of $f(\mathbf{B}^d)$ can be derived as

$$\nabla f = \mathbf{M}^d - \boldsymbol{\rho}^d - \frac{1}{\sigma_B^2}\mathbf{B}^d.$$

To compute the Hessian, it is easier to define the gradient ∇f as a vector, instead of a matrix, and hence we stack the columns of $\mathbf{B}^d$ into $\boldsymbol{\beta}^d$, i.e., $\boldsymbol{\beta}^d = \mathbf{B}^d(:)$ for avid Matlab users. The Hessian matrix can now be readily computed taking the derivatives of the gradient, yielding

$$\begin{aligned}\nabla\nabla f &= -\frac{1}{\sigma_B^2}\mathbf{I}_{RK} + \nabla\nabla\log p(\mathbf{x}^d|\boldsymbol{\beta}^d, \mathbf{Z}) \\ &= -\frac{1}{\sigma_B^2}\mathbf{I}_{RK} - \sum_{n=1}^{N}\left(\text{diag}(\boldsymbol{\pi}_n^d) - (\boldsymbol{\pi}_n^d)^\top\boldsymbol{\pi}_n^d\right)\otimes(\mathbf{z}_n^\top\mathbf{z}_n),\end{aligned}$$

where $\text{diag}(\boldsymbol{\pi}_n^d)$ is a diagonal matrix with the values of the vector $\boldsymbol{\pi}_n^d = \left[\pi_{n1}^d, \pi_{n2}^d, \ldots, \pi_{nR}^d\right]$ as its diagonal elements.

Finally, note that, since $p(\mathbf{x}^d|\boldsymbol{\beta}^d, \mathbf{Z})$ is a log-concave function of $\boldsymbol{\beta}^d$ [13, p. 87], $-\nabla\nabla f$ is a positive definite matrix, which guarantees that the maximum of $f(\boldsymbol{\beta}^d)$ is unique.

Appendix B

Nested EP: Inner loop

The inner loop is an EP method that approximates by a Gaussian the tilted distribution $\widehat{p}_n(\boldsymbol{\beta}^d)$, which can be expressed as

$$\begin{aligned}\widehat{p}_n(\boldsymbol{\beta}^d) &= \frac{1}{\widehat{Z}_n} q_{\neg n}(\boldsymbol{\beta}^d) t_n^d(\boldsymbol{\beta}^d) \\ &= \frac{1}{\widehat{Z}_n} \mathcal{N}(\boldsymbol{\beta}^d | \boldsymbol{\Pi}_{\neg n}^{-1} \boldsymbol{\lambda}_{\neg n}, \boldsymbol{\Pi}_{\neg n}^{-1}) \times \int \mathcal{N}(u_{nd}|0,1) \left(\prod_{\substack{r=1 \\ r \neq x_n^d}}^{R} \Phi\left(u_{nd} + \mathbf{z}_n (\mathbf{b}_{\cdot x_n^d}^d - \mathbf{b}_{\cdot r}^d) \right) \right) du_{nd}.\end{aligned} \tag{B.1}$$

Removing the marginalization with respect to the auxiliary variable u_{nd} and defining $\boldsymbol{\beta}_{\mathrm{I}}^d$ as the vector compound of $\boldsymbol{\beta}^d$ and u_{nd}, namely, $\boldsymbol{\beta}_{\mathrm{I}}^d = \left[(\boldsymbol{\beta}^d)^\top, u_{nd} \right]^\top$, we have the augmented tilted distribution

$$\widehat{p}_n(\boldsymbol{\beta}_{\mathrm{I}}^d) = \frac{1}{\widehat{Z}_n} \mathcal{N}(\boldsymbol{\beta}_{\mathrm{I}}^d | \boldsymbol{\Pi}_{\mathrm{I}_n}^{-1} \boldsymbol{\lambda}_{\mathrm{I}_n}, \boldsymbol{\Pi}_{\mathrm{I}_n}^{-1}) \prod_{\substack{r=1 \\ r \neq x_n^d}}^{R} \Phi\left((\mathbf{h}_{nr}^d)^\top \boldsymbol{\beta}_{\mathrm{I}}^d \right), \tag{B.2}$$

where we have defined $\boldsymbol{\Pi}_{\mathrm{I}n}$ as a block-diagonal matrix formed from $\boldsymbol{\Pi}_{\neg n}$ and 1, $\boldsymbol{\lambda}_{\mathrm{I}_n} = \left[\boldsymbol{\lambda}_{\neg n}^\top, 0 \right]^\top$, and $\mathbf{h}_{nr}^d = \left[(\mathbf{e}_{x_n^d} - \mathbf{e}_r)^\top \otimes \mathbf{z}_n,\ 1 \right]^\top$. Here, '$\otimes$' denotes the Kronecker product, and $\mathbf{e}_r$ is the r-th unit (column) vector of the R-dimensional standard basis. Note that we use the subscript 'I' to denote the augmented variables that account for both $\boldsymbol{\beta}^d$ and u_{nd}. The normalization term $\widehat{Z}_n$ is the same for $\widehat{p}_n(\boldsymbol{\beta}^d)$ and for the augmented distribution $\widehat{p}_n(\boldsymbol{\beta}_{\mathrm{I}}^d)$,

and it is defined as

$$\widehat{Z}_n = \int q_{\neg n}(\boldsymbol{\beta}^d)\mathcal{N}(u_{nd}|0,1)\prod_{r\neq x_n^d}\Phi((\mathbf{h}_{nr}^d)^\top\boldsymbol{\beta}_{\mathrm{I}}^d)d\boldsymbol{\beta}_{\mathrm{I}}^d. \tag{B.3}$$

Due to the multinomial probit model, Eq. B.2 contains a product of intractable functions of the scalar variables $s_r = (\mathbf{h}_{nr}^d)^\top\boldsymbol{\beta}_{\mathrm{I}}^d$, allowing us to apply a new inner EP loop, which is simpler than the outer loop since it only involves scalar operations. Hence, the augmented distribution in (B.2) can be approximated by replacing each intractable term $\Phi(s_r)$ with a scaled univariate Gaussian site function with natural parameters $\widetilde{\alpha}_{nr}$ and $\widetilde{\beta}_{nr}$, resulting in the approximate distribution

$$\begin{aligned} q_{\mathrm{I}_n}(\boldsymbol{\beta}_{\mathrm{I}}^d) &= \frac{1}{C_{\mathrm{I}_n}}\mathcal{N}(\boldsymbol{\beta}_{\mathrm{I}}^d|\boldsymbol{\Pi}_{\mathrm{I}_n}^{-1}\boldsymbol{\lambda}_{\mathrm{I}_n},\boldsymbol{\Pi}_{\mathrm{I}_n}^{-1})\prod_{\substack{r=1\\ r\neq x_n^d}}^{R}\widetilde{C}_{nr}\mathcal{N}(s_r|\widetilde{\alpha}_{nr}^{-1}\widetilde{\beta}_{nr},\widetilde{\alpha}_{nr}^{-1}) \\ &= \mathcal{N}(\boldsymbol{\beta}_{\mathrm{I}}^d|\widetilde{\boldsymbol{\Pi}}_{\mathrm{I}_n}^{-1}\widetilde{\boldsymbol{\lambda}}_{\mathrm{I}_n},\widetilde{\boldsymbol{\Pi}}_{\mathrm{I}_n}^{-1}), \end{aligned} \tag{B.4}$$

where the normalization constant C_{I_n} approximates $\widehat{Z}_n$.

We start from $q_{nr}(s_r) = \mathcal{N}(s_r|m_{nr},v_{nr})$, being $m_{nr} = (\mathbf{h}_{nr}^d)^\top\widetilde{\boldsymbol{\Pi}}_{\mathrm{I}_n}^{-1}\widetilde{\boldsymbol{\lambda}}_{\mathrm{I}_n}$ and $v_{nr} = (\mathbf{h}_{nr}^d)^\top\widetilde{\boldsymbol{\Pi}}_{\mathrm{I}_n}^{-1}\mathbf{h}_{nr}^d$. Then, the cavity distribution $q_{n\neg r}(s_r)$ can be written as

$$q_{n\neg r}(s_r) = \mathcal{N}(s_r|m_{n\neg r},v_{n\neg r}), \tag{B.5}$$

which has mean $m_{n\neg r} = v_{n\neg r}(m_{nr}/v_{nr} - \widetilde{\beta}_{nr})$ and variance $v_{n\neg r} = (1/v_{nr} + \widetilde{\alpha}_{nr})^{-1}$. The tilted distribution (including one true site),

$$\widehat{f}_{nr}(s_r) = \frac{1}{\widehat{C}_{nr}}q_{n\neg r}(s_r)\Phi(s_r), \tag{B.6}$$

has mean $\widehat{m}_{nr} = m_{n\neg r} + \rho_{nr}v_{n\neg r}$, variance $\widehat{v}_{nr} = v_{n\neg r} - v_{n\neg r}^2\left(\rho_{nr}^2 + \rho_{nr}\frac{m_{n\neg r}}{1+v_{n\neg r}}\right)$ and normalization constant $\widehat{C}_{nr} = \Phi\left(\frac{m_{n\neg r}}{\sqrt{1+v_{n\neg r}}}\right)$, being

$$\rho_{nr} = \frac{\mathcal{N}\left(\frac{m_{n\neg r}}{\sqrt{1+v_{n\neg r}}}|0,1\right)}{\Phi\left(\frac{m_{n\neg r}}{\sqrt{1+v_{n\neg r}}}\right)\sqrt{1+v_{n\neg r}}}. \tag{B.7}$$

Finally, the site updates are computed as $\widetilde{\alpha}_{nr} = 1/\widehat{v}_{nr} - 1/v_{n\neg r}$ and $\widetilde{\beta}_{nr} = \widehat{m}_{nr}/\widehat{v}_{nr} - m_{n\neg r}/v_{n\neg r}$. Again, a damping factor of η_{I} can be used in this step. In this case, the site updates can be obtained in parallel for

the different values of r, afterwards recomputing the natural parameters of $q_{\mathrm{I}_n}(\boldsymbol{\beta}_{\mathrm{I}}^d)$ as $\widetilde{\boldsymbol{\Pi}}_{\mathrm{I}_n} = \boldsymbol{\Pi}_{\mathrm{I}_n} + \sum_{r \neq x_n^d} \widetilde{\alpha}_{nr} \mathbf{h}_{nr}^d (\mathbf{h}_{nr}^d)^\top$ and $\widetilde{\boldsymbol{\lambda}}_{\mathrm{I}_n} = \boldsymbol{\lambda}_{\mathrm{I}_n} + \sum_{r \neq x_n^d} \widetilde{\beta}_{nr} \mathbf{h}_{nr}^d$.

The constants C_{I_n} (which approximates $\widehat{Z}_n$ in Eq. B.2) and $\widetilde{C}_{nr}$ in (B.4) can be computed after meeting the stopping criterion as

$$\begin{aligned} \log C_{\mathrm{I}_n} = & \sum_{\substack{r=1 \\ r \neq x_n^d}}^{R} \left(\log \widetilde{C}_{nr} + \frac{1}{2} \log(\widetilde{\alpha}_{nr}) \right) \\ & + \frac{1}{2} \log(|\boldsymbol{\Pi}_{\mathrm{I}_n}| - |\widetilde{\boldsymbol{\Pi}}_{\mathrm{I}_n}|) + \frac{1}{2} \left(\widetilde{\boldsymbol{\lambda}}_{\mathrm{I}_n}^\top \widetilde{\boldsymbol{\Pi}}_{\mathrm{I}_n}^{-1} \widetilde{\boldsymbol{\lambda}}_{\mathrm{I}_n} - \boldsymbol{\lambda}_{\mathrm{I}_n}^\top \boldsymbol{\Pi}_{\mathrm{I}_n}^{-1} \boldsymbol{\lambda}_{\mathrm{I}_n} \right), \end{aligned} \tag{B.8}$$

and

$$\log \widetilde{C}_{nr} = \log \widehat{C}_{nr} + \frac{1}{2} \log(v_{n\neg r} + 1/\widetilde{\alpha}_{nr}) + \frac{1}{2} \left(\frac{m_{n\neg r}^2}{v_{n\neg r}} - \frac{\left(\dfrac{m_{n\neg r}}{v_{n\neg r}} + \widetilde{\beta}_{nr} \right)^2}{1/v_{n\neg r} + \widetilde{\alpha}_{nr}} \right). \tag{B.9}$$

Matrices $\widehat{\boldsymbol{\Pi}}_n$ and $\widehat{\boldsymbol{\lambda}}_n$ of the outer loop can be obtained from $\widetilde{\boldsymbol{\Pi}}_{\mathrm{I}_n}$ and $\widetilde{\boldsymbol{\lambda}}_{\mathrm{I}_n}$ after removing the effects of the auxiliary variable u_{nd}.

Appendix C

Variational Inference Derivation

C.1 Lower Bound Derivation

In this chapter we derive the lower bound $\mathcal{L}(\mathcal{H}, \mathcal{H}_q)$ on the evidence $p(\mathbf{X}|\mathcal{H})$. From Eq. (4.23),

$$\begin{aligned}\log p(\mathbf{X}|\mathcal{H}) &= \mathbb{E}_q\left[\log p(\Psi, \mathbf{X}|\mathcal{H})\right] + H[q] + D_{KL}(q||p) \\ &\geqslant \mathbb{E}_q\left[\log p(\Psi, \mathbf{X}|\mathcal{H})\right] + H[q].\end{aligned}$$

The expectation $\mathbb{E}_q\left[\log p(\Psi, \mathbf{X}|\mathcal{H})\right]$ can be derived as

$$\begin{aligned}&\mathbb{E}_q\left[\log p(\Psi, \mathbf{X}|\mathcal{H})\right] \\ &= \sum_{k=1}^{K} \underbrace{\mathbb{E}_q\left[\log p(v_k|\alpha)\right]}_{1} + \sum_{d=1}^{D}\sum_{k=1}^{K} \underbrace{\mathbb{E}_q\left[\log p(\mathbf{b}_{k\cdot}^d|\sigma_B^2)\right]}_{2} + \sum_{d=1}^{D} \underbrace{\mathbb{E}_q\left[\log p(\mathbf{b}_0^d|\sigma_B^2)\right]}_{3} \\ &\quad + \sum_{k=1}^{K}\sum_{n=1}^{N} \underbrace{\mathbb{E}_q\left[\log p(z_{nk}|\{v_i\}_{i=1}^k)\right]}_{4} + \sum_{n=1}^{N}\sum_{d=1}^{D} \underbrace{\mathbb{E}_q\left[\log p(x_{nd}|\mathbf{z}_n, \mathbf{B}^d, \mathbf{b}_0^d)\right]}_{5},\end{aligned} \tag{C.1}$$

where each term can be computed as shown below:

1. For the Beta distribution over v_k,

$$\mathbb{E}_q\left[\log p(v_k|\alpha)\right] = \log(\alpha) + (\alpha - 1)\left[\psi(\tau_{k1}) - \psi(\tau_{k1} + \tau_{k2})\right].$$

2. For the Gaussian distribution over vectors $\mathbf{b}_{k\cdot}^d$,

$$\mathbb{E}_q\left[\log p(\mathbf{b}_{k\cdot}^d|\sigma_B^2)\right] = -\frac{R}{2}\log(2\pi\sigma_B^2) - \frac{1}{2\sigma_B^2}\left(\sum_{r=1}^{R}(\phi_{kr}^d)^2 + \sum_{r=1}^{R}\left(\sigma_{kr}^d\right)^2\right).$$

3. For the Gaussian distribution over $\mathbf{b}_0^d$,

$$\mathbb{E}_q\left[\log p(\mathbf{b}_0^d|\sigma_B^2)\right] = -\frac{R}{2}\log(2\pi\sigma_B^2) - \frac{1}{2\sigma_B^2}\left(\sum_{r=1}^{R}(\phi_{0r}^d)^2 + \sum_{r=1}^{R}(\sigma_{0r}^d)^2\right).$$

4. For the feature assignments, which are Bernoulli distributed given the feature probabilities, we have

$$\begin{aligned}
&\mathbb{E}_q\left[\log p(z_{nk}|\{v_i\}_{i=1}^k)\right] \\
&= (1-\nu_{nk})\mathbb{E}_q\left[\log\left(1-\prod_{i=1}^{k}v_i\right)\right] + \nu_{nk}\sum_{i=1}^{k}\left[\psi(\tau_{i1}) - \psi(\tau_{i1}+\tau_{i2})\right],
\end{aligned}$$

where the expectation $\mathbb{E}_q\left[\log\left(1-\prod_{i=1}^{k}v_i\right)\right]$ has no closed-form solution. We can instead lower bound it by using the multinomial approach [19]. Under this approach, we introduce an auxiliary multinomial distribution $\boldsymbol{\lambda}_k = [\lambda_{k1},\ldots,\lambda_{kk}]$ in the expectation and apply Jensen's inequality, yielding

$$\begin{aligned}
&\mathbb{E}_q\left[\log\left(1-\prod_{i=1}^{k}v_i\right)\right] \\
&\geq \sum_{m=1}^{k}\lambda_{km}\psi(\tau_{m2}) + \sum_{m=1}^{k-1}\left(\sum_{n=m+1}^{k}\lambda_{kn}\right)\psi(\tau_{m1}) \\
&\quad - \sum_{m=1}^{k}\left(\sum_{n=m}^{k}\lambda_{kn}\right)\psi(\tau_{m1}+\tau_{m2}) - \sum_{m=1}^{k}\lambda_{km}\log(\lambda_{km}),
\end{aligned}$$

which holds for any distribution represented by the probabilities

$\lambda_{k1}, \ldots, \lambda_{kk}$, for $1 \leq k \leq K$. Then,

$$\begin{aligned}
&\mathbb{E}_q \left[\log p(z_{nk} | \{v_i\}_{i=1}^k) \right] \\
&\geq (1 - \nu_{nk}) \left[\sum_{m=1}^{k} \lambda_{km} \psi(\tau_{m2}) + \sum_{m=1}^{k-1} \left(\sum_{n=m+1}^{k} \lambda_{kn} \right) \psi(\tau_{m1}) \right. \\
&\quad \left. - \sum_{m=1}^{k} \left(\sum_{n=m}^{k} \lambda_{kn} \right) \psi(\tau_{m1} + \tau_{m2}) - \sum_{m=1}^{k} \lambda_{km} \log(\lambda_{km}) \right] \\
&\quad + \nu_{nk} \sum_{i=1}^{k} \left[\psi(\tau_{i1}) - \psi(\tau_{i1} + \tau_{i2}) \right].
\end{aligned}$$

5. For the likelihood term, we can write

$$\begin{aligned}
&\mathbb{E}_q \left[\log p(x_{nd} | \mathbf{z}_n, \mathbf{B}^d, \mathbf{b}_0^d) \right] \\
&= \phi_{0x_{nd}}^d + \sum_{k=1}^{K} \nu_{nk} \phi_{kx_{nd}}^d - \mathbb{E}_q \left[\log \left(\sum_{r=1}^{R} \exp(\mathbf{z}_n \mathbf{b}_{\cdot r}^d + b_{0r}^d) \right) \right],
\end{aligned}$$

where the logarithm can be upper bounded by its first-order Taylor series expansion around the auxiliary variable ξ_{nd}^{-1} (for $n = 1, \ldots, N$ and $d = 1, \ldots, D$) [11, 12], yielding

$$\begin{aligned}
&\log \left(\sum_{r=1}^{R} \exp(\mathbf{z}_n \mathbf{b}_{\cdot r}^d + b_{0r}^d) \right) \\
&\leq \xi_{nd} \left(\sum_{r=1}^{R} \exp(\mathbf{z}_n \mathbf{b}_{\cdot r}^d + b_{0r}^d) \right) - \log(\xi_{nd}) - 1.
\end{aligned}$$

The main advantage of this bound lies on the fact that it allows us to compute the expectation of the bound for the Gaussian distribution, since it involves the moment generating functions of the distributions $q(\mathbf{b}_{\cdot r}^d)$ and $q(b_{0r}^d)$. Then, we can lower bound the likelihood term as

$$\begin{aligned}
&\mathbb{E}_q \left[\log p(x_{nd} | \mathbf{z}_n, \mathbf{B}^d, \mathbf{b}_0^d) \right] \\
&\geq \phi_{0x_{nd}}^d + \sum_{k=1}^{K} \nu_{nk} \phi_{kx_{nd}}^d + \log(\xi_{nd}) + 1 - \xi_{nd} \sum_{r=1}^{R} \left[\exp\left(\phi_{0r}^d + \frac{1}{2} (\sigma_{0r}^d)^2 \right) \right. \\
&\quad \left. \times \prod_{k=1}^{K} \left(1 - \nu_{nk} + \nu_{nk} \exp\left(\phi_{kr}^d + \frac{1}{2} (\sigma_{kr}^d)^2 \right) \right) \right].
\end{aligned}$$

Substituting the previous results in (C.1), we obtain

$$
\begin{aligned}
&\mathbb{E}_q\left[\log p(\Psi, \mathbf{X}|\mathcal{H})\right] \\
&\geq \sum_{k=1}^{K}\left[\log(\alpha) + (\alpha-1)\left(\psi(\tau_{k1}) - \psi(\tau_{k1}+\tau_{k2})\right)\right] \\
&\quad - \frac{R(K+1)D}{2}\log(2\pi\sigma_B^2) - \frac{1}{2\sigma_B^2}\sum_{k=0}^{K}\sum_{d=1}^{D}\sum_{r=1}^{R}\left((\phi_{kr}^d)^2 + (\sigma_{kr}^d)^2\right) \\
&\quad + \sum_{n=1}^{N}\sum_{k=1}^{K}\Bigg[\nu_{nk}\sum_{i=1}^{k}\left[\psi(\tau_{i1}) - \psi(\tau_{i1}+\tau_{i2})\right] \\
&\qquad\qquad + (1-\nu_{nk})\left(\sum_{m=1}^{k}\lambda_{km}\psi(\tau_{m2}) + \sum_{m=1}^{k-1}\left(\sum_{n=m+1}^{k}\lambda_{kn}\right)\psi(\tau_{m1})\right. \\
&\qquad\qquad \left. - \sum_{m=1}^{k}\left(\sum_{n=m}^{k}\lambda_{kn}\right)\psi(\tau_{m1}+\tau_{m2}) - \sum_{m=1}^{k}\lambda_{km}\log(\lambda_{km})\right)\Bigg] \\
&\quad + \sum_{n=1}^{N}\sum_{d=1}^{D}\Bigg[\phi_{0x_{nd}}^d + \sum_{k=1}^{K}\nu_{nk}\phi_{kx_{nd}}^d + \log(\xi_{nd}) + 1 \\
&\quad - \xi_{nd}\sum_{r=1}^{R}\left[\exp\left(\phi_{0r}^d + \frac{1}{2}(\sigma_{0r}^d)^2\right)\prod_{k=1}^{K}\left(1-\nu_{nk}+\nu_{nk}\exp\left(\phi_{kr}^d + \frac{1}{2}(\sigma_{kr}^d)^2\right)\right)\right]\Bigg].
\end{aligned}
$$

Additionally, the entropy of the distribution $q(\Psi)$ is given by

$$
\begin{aligned}
H[q] &= \mathbb{E}_q\left[\log q(\Psi)\right] \\
&= \sum_{k=1}^{K} \mathbb{E}_q\left[\log q(v_k|\tau_{k1},\tau_{k2})\right] + \sum_{d=1}^{D}\sum_{r=1}^{R}\sum_{k=0}^{K} \mathbb{E}_q\left[\log q(b_{kr}^d|\phi_{kr}^d,(\sigma_{kr}^d)^2)\right] \\
&\quad + \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}_q\left[\log q(z_{nk}|\nu_{nk})\right] \\
&= \sum_{k=1}^{K}\Big[\log\left(\frac{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}{\Gamma(\tau_{k1}+\tau_{k2})}\right) - (\tau_{k1}-1)\psi(\tau_{k1}) - (\tau_{k2}-1)\psi(\tau_{k2}) \\
&\qquad\quad + (\tau_{k1}+\tau_{k2}-2)\psi(\tau_{k1}+\tau_{k2})\Big] \\
&\quad + \sum_{d=1}^{D}\sum_{r=1}^{R}\sum_{k=0}^{K} \frac{1}{2}\log(2\pi e(\sigma_{kr}^d)^2) \\
&\quad + \sum_{n=1}^{N}\sum_{k=1}^{K}\left[-\nu_{nk}\log(\nu_{nk}) - (1-\nu_{nk})\log(1-\nu_{nk})\right].
\end{aligned}
$$

Finally, we obtain the lower bound on the evidence $p(\mathbf{X}|\mathcal{H})$ as

$$
\begin{aligned}
&\log p(\mathbf{X}|\mathcal{H}) \geq \mathbb{E}_q\left[\log p(\Psi, \mathbf{X}|\mathcal{H})\right] + H[q] \\
&\geq \sum_{k=1}^{K}\left[\log(\alpha) + (\alpha-1)\left(\psi(\tau_{k1}) - \psi(\tau_{k1}+\tau_{k2})\right)\right] \\
&\quad - \frac{R(K+1)D}{2}\log(2\pi\sigma_B^2) - \frac{1}{2\sigma_B^2}\sum_{k=0}^{K}\sum_{d=1}^{D}\sum_{r=1}^{R}\left((\phi_{kr}^d)^2 + (\sigma_{kr}^d)^2\right) \\
&\quad + \sum_{n=1}^{N}\sum_{k=1}^{K}\Bigg[\nu_{nk}\sum_{i=1}^{k}\left[\psi(\tau_{i1}) - \psi(\tau_{i1}+\tau_{i2})\right] \\
&\qquad\qquad + (1-\nu_{nk})\left(\sum_{m=1}^{k}\lambda_{km}\psi(\tau_{m2}) + \sum_{m=1}^{k-1}\left(\sum_{n=m+1}^{k}\lambda_{kn}\right)\psi(\tau_{m1})\right. \\
&\qquad\qquad \left. - \sum_{m=1}^{k}\left(\sum_{n=m}^{k}\lambda_{kn}\right)\psi(\tau_{m1}+\tau_{m2}) - \sum_{m=1}^{k}\lambda_{km}\log(\lambda_{km})\right)\Bigg] \\
&\quad + \sum_{n=1}^{N}\sum_{d=1}^{D}\Bigg[\phi_{0x_{nd}}^d + \sum_{k=1}^{K}\nu_{nk}\phi_{kx_{nd}}^d + \log(\xi_{nd}) + 1 \\
&\quad -\xi_{nd}\sum_{r=1}^{R}\left[\exp\left(\phi_{0r}^d + \frac{1}{2}(\sigma_{0r}^d)^2\right)\prod_{k=1}^{K}\left(1-\nu_{nk}+\nu_{nk}\exp\left(\phi_{kr}^d + \frac{1}{2}(\sigma_{kr}^d)^2\right)\right)\right]\Bigg] \\
&\quad + \sum_{k=1}^{K}\left[\log\left(\frac{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}{\Gamma(\tau_{k1}+\tau_{k2})}\right) - (\tau_{k1}-1)\psi(\tau_{k1}) - (\tau_{k2}-1)\psi(\tau_{k2}) + (\tau_{k1}+\tau_{k2}-2)\psi(\tau_{k1}+\tau_{k2})\right] \\
&\quad + \sum_{d=1}^{D}\sum_{r=1}^{R}\sum_{k=0}^{K}\frac{1}{2}\log(2\pi e(\sigma_{kr}^d)^2) + \sum_{n=1}^{N}\sum_{k=1}^{K}\left[-\nu_{nk}\log(\nu_{nk}) - (1-\nu_{nk})\log(1-\nu_{nk})\right] \\
&= \mathcal{L}(\mathcal{H}, \mathcal{H}_q),
\end{aligned}
$$

where $\mathcal{H}_q = \{\tau_{k1}, \tau_{k2}, \lambda_{km}, \xi_{nd}, \nu_{nk}, \phi_{kr}^d, \phi_{0r}^d, (\sigma_{kr}^d)^2, (\sigma_{0r}^d)^2\}$ (for $k = 1, \ldots, K$, $m = 1, \ldots, k$, $d = 1, \ldots, D$, and $n = 1, \ldots, N$) represents the set of the variational parameters.

C.2 Derivatives for Newton's Method

- For the parameters of the Gaussian distribution $q(b_{kr}^d|\phi_{kr}^d,(\sigma_{kr}^d)^2)$ for $k=1,\ldots,K$,

$$\frac{\partial}{\partial\phi_{kr}^d}\mathcal{L}(\mathcal{H},\mathcal{H}_q)$$
$$=-\frac{1}{\sigma_B^2}\phi_{kr}^d+\sum_{n=1}^{N}\left[\nu_{nk}\delta(x_{nd}=r)-\nu_{nk}\xi_{nd}\exp\left(\phi_{0r}^d+\frac{1}{2}(\sigma_{0r}^d)^2\right)\exp\left(\phi_{kr}^d+\frac{1}{2}(\sigma_{kr}^d)^2\right)\right.$$
$$\left.\times\prod_{k'\neq k}\left(1-\nu_{nk'}+\nu_{nk'}\exp\left(\phi_{k'r}^d+\frac{1}{2}(\sigma_{k'r}^d)^2\right)\right)\right].$$

$$\frac{\partial^2}{\partial(\phi_{kr}^d)^2}\mathcal{L}(\mathcal{H},\mathcal{H}_q)$$
$$=-\frac{1}{\sigma_B^2}-\sum_{n=1}^{N}\left[\nu_{nk}\xi_{nd}\exp\left(\phi_{0r}^d+\frac{1}{2}(\sigma_{0r}^d)^2\right)\exp\left(\phi_{kr}^d+\frac{1}{2}(\sigma_{kr}^d)^2\right)\right.$$
$$\left.\times\prod_{k'\neq k}\left(1-\nu_{nk'}+\nu_{nk'}\exp\left(\phi_{k'r}^d+\frac{1}{2}(\sigma_{k'r}^d)^2\right)\right)\right].$$

$$\frac{\partial}{\partial(\sigma_{kr}^d)^2}\mathcal{L}(\mathcal{H},\mathcal{H}_q)$$
$$=-\frac{1}{2\sigma_B^2}+\frac{1}{2}(\sigma_{kr}^d)^{-2}-\frac{1}{2}\sum_{n=1}^{N}\left[\nu_{nk}\xi_{nd}\exp\left(\phi_{0r}^d+\frac{1}{2}(\sigma_{0r}^d)^2\right)\exp\left(\phi_{kr}^d+\frac{1}{2}(\sigma_{kr}^d)^2\right)\right.$$
$$\left.\times\prod_{k'\neq k}\left(1-\nu_{nk'}+\nu_{nk'}\exp\left(\phi_{k'r}^d+\frac{1}{2}(\sigma_{k'r}^d)^2\right)\right)\right].$$

$$\frac{\partial^2}{(\partial(\sigma_{kr}^d)^2)^2}\mathcal{L}(\mathcal{H},\mathcal{H}_q)$$
$$=-\frac{1}{2}(\sigma_{kr}^d)^{-4}-\frac{1}{4}\sum_{n=1}^{N}\left[\nu_{nk}\xi_{nd}\exp\left(\phi_{0r}^d+\frac{1}{2}(\sigma_{0r}^d)^2\right)\exp\left(\phi_{kr}^d+\frac{1}{2}(\sigma_{kr}^d)^2\right)\right.$$
$$\left.\times\prod_{k'\neq k}\left(1-\nu_{nk'}+\nu_{nk'}\exp\left(\phi_{k'r}^d+\frac{1}{2}(\sigma_{k'r}^d)^2\right)\right)\right].$$

- For the parameters of the Gaussian distribution $q(b_{0r}^d|\phi_{0r}^d,(\sigma_{0r}^d)^2)$,

$$\frac{\partial}{\partial\phi_{0r}^d}\mathcal{L}(\mathcal{H},\mathcal{H}_q)=-\frac{1}{\sigma_B^2}\phi_{0r}^d+\sum_{n=1}^{N}\Bigg[\delta(x_{nd}=r)$$
$$-\xi_{nd}\exp\left(\phi_{0r}^d+\frac{1}{2}(\sigma_{0r}^d)^2\right)\prod_{k=1}^{K}\left(1-\nu_{nk}+\nu_{nk}\exp\left(\phi_{kr}^d+\frac{1}{2}(\sigma_{kr}^d)^2\right)\right)\Bigg].$$

$$\frac{\partial^2}{(\partial\phi_{0r}^d)^2}\mathcal{L}(\mathcal{H},\mathcal{H}_q)$$
$$=-\frac{1}{\sigma_B^2}-\sum_{n=1}^{N}\left[\xi_{nd}\exp\left(\phi_{0r}^d+\frac{1}{2}(\sigma_{0r}^d)^2\right)\prod_{k=1}^{K}\left(1-\nu_{nk}+\nu_{nk}\exp\left(\phi_{kr}^d+\frac{1}{2}(\sigma_{kr}^d)^2\right)\right)\right].$$

$$\frac{\partial}{\partial(\sigma_{0r}^d)^2}\mathcal{L}(\mathcal{H},\mathcal{H}_q)=-\frac{1}{2\sigma_B^2}+\frac{1}{2}(\sigma_{0r}^d)^{-2}$$
$$-\frac{1}{2}\sum_{n=1}^{N}\left[\xi_{nd}\exp\left(\phi_{0r}^d+\frac{1}{2}(\sigma_{0r}^d)^2\right)\prod_{k=1}^{K}\left(1-\nu_{nk}+\nu_{nk}\exp\left(\phi_{kr}^d+\frac{1}{2}(\sigma_{kr}^d)^2\right)\right)\right].$$

$$\frac{\partial^2}{(\partial(\sigma_{0r}^d)^2)^2}\mathcal{L}(\mathcal{H},\mathcal{H}_q)=-\frac{1}{2}(\sigma_{0r}^d)^{-4}$$
$$-\frac{1}{4}\sum_{n=1}^{N}\left[\xi_{nd}\exp\left(\phi_{0r}^d+\frac{1}{2}(\sigma_{0r}^d)^2\right)\prod_{k=1}^{K}\left(1-\nu_{nk}+\nu_{nk}\exp\left(\phi_{kr}^d+\frac{1}{2}(\sigma_{kr}^d)^2\right)\right)\right].$$

Appendix D

NESARC Survey

In this chapter, we show Sections 10 and 11 of the NESARC survey. These section contain the questions necessary to diagnose the seven personality disorders studied in Chapter 8, i.e., avoidant, dependent, obsessive-compulsive, paranoid, schizoid, histrionic and antisocial.

Section 10 - USUAL FEELINGS AND ACTIONS

Statement S

The questions I'm going to ask you now are about how you have felt or acted MOST of the time throughout your life regardless of the situation or whom you were with. Do NOT include times when you weren't yourself or when you acted differently than usual because you were depressed or hyper, anxious or nervous or drinking heavily, using medicines or drugs or experiencing their bad aftereffects, or times when you were physically ill.

1a. Most of the time throughout your life, regardless of the situation or whom you were with. . . (*Repeat phrase frequently)*		**b. Did this ever trouble you or cause problems at work or school, or with your family or other people?**
(1) Have you avoided jobs or tasks that dealt with a lot of people?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(2) Do you avoid getting involved with people unless you are certain they will like you?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(3) Do you find it hard to be "open" even with people you are close to?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(4) Do you often worry about being criticized or rejected in social situations?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(5) Do you believe that you're not as good, as smart, or as attractive as most other people?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(6) Are you usually quiet or do you have very little to say when you meet new people because you believe they are better than you are?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(7) Are you afraid of trying new things or doing things outside your usual routine because you're afraid of being embarrassed?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(8) Do you need a lot of advice or reassurance from others before you can make everyday decisions-like what to wear or what to order in a restaurant?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(9) Do you depend on other people to handle important areas in your life such as finances, child care, or living arrangements?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(10) Do you find it hard to disagree with people even when you think they are wrong because you fear losing their support or approval?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(11) Do you find it hard to start or work on tasks when there is no one to help you?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(12) Have you often volunteered to do things even if they are unpleasant in order to get others to like you?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(13) Do you usually feel uncomfortable when you are by yourself because you are afraid you can't take care of yourself?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(14) When a close relationship ends, do you feel you immediately have to find someone else to take care of you?	1 ☐ Yes → 2 ☐ No - *Go to next experience, page 110*	1 ☐ Yes 2 ☐ No

Section 10 - USUAL FEELINGS AND ACTIONS (Continued)

1a. Most of the time throughout your life, regardless of the situation or whom you were with. . . (*Repeat phrase frequently)*		**b.** Did this ever trouble you or cause problems at work or school, or with your family or other people?
(15) Have you worried a lot about being left alone to take care of yourself?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(16) Are you the kind of person who focuses on details, order and organization or likes to make lists and schedules?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(17) Do you sometimes get so caught up with details, schedules or organization that you lose sight of what you wanted to accomplish?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(18) Do you have trouble finishing jobs because you spend so much time trying to get things exactly right?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(19) Do you or other people feel that you are so devoted to work or school that you have no time left for anyone else or for just having fun?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(20) Do other people think you have unreasonably high standards and morals about what is right and what is wrong?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(21) Do you have trouble throwing out worn-out or worthless things even if they don't have sentimental value?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(22) Is it hard for you to let other people help you if they don't agree to do things exactly the way you want?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(23) Is it hard for you to spend money on yourself and other people even when you have enough?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(24) Are you often so sure you are right that it doesn't matter what other people say?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(25) Have other people told you that you are stubborn or rigid?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(26) Do you often have to keep an eye out to keep people from using you, hurting you or lying to you?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(27) Do you spend a lot of time wondering if you can trust your friends or the people you work with?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(28) Do you find that it is best not to let other people know much about you because they will use it against you?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(29) Do you often detect hidden threats or insults in things people say or do?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(30) Are you the kind of person who takes a long time to forgive people who have insulted or slighted you?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience, page 111*	1 ☐ Yes 2 ☐ No

Section 10 - USUAL FEELINGS AND ACTIONS (Continued)		
1a. Most of the time throughout your life, regardless of the situation or whom you were with . . . (*Repeat phrase frequently)*		**b. Did this ever trouble you or cause problems at work or school, or with your family or other people?**
(31) Have there been many people you can't forgive because they did or said something to you a long time ago?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(32) Do you often get angry or lash out when someone criticizes or insults you in some way?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(33) Have you OFTEN suspected that your spouse or partner has been unfaithful?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(35) When you are around people, do you often feel that you are being watched or stared at?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(43) Are there very few people that you're really close to outside of your immediate family?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(45) Would you be just as happy without having any close relationships?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(46) Do you take little pleasure in being with other people?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(47) Have you almost always preferred to do things alone rather than with other people?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(48) Could you be content without ever being sexually involved with anyone?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(49) Do you rarely show much emotion?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(50) Are there really very few things that give you pleasure?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(51) Do you rarely react to praise or criticism?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(52) Are you the sort of person who doesn't care about what people think of you?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(53) Do you find that nothing makes you very happy or very sad?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(54) Do you like to be the center of attention?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(55) Do your feelings often change very suddenly or unexpectedly, sometimes for no reason?	1 ☐ Yes ⟶ 2 ☐ No - *Go to next experience, page 112*	1 ☐ Yes 2 ☐ No

Section 10 - USUAL FEELINGS AND ACTIONS (Continued)

1a. Most of the time throughout your life, regardless of the situation or whom you were with . . . (*Repeat phrase frequently)*		b. Did this ever trouble you or cause problems at work or school, or with your family or other people?
(56) Do you feel uncomfortable if you are not the center of attention?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(57) Have you ever discovered that people aren't as close to you as you thought they were?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(58) Do you flirt a lot?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(59) Do you display your emotions in obvious or dramatic ways so that people always know how you feel?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(60) Do you often find yourself "coming on" to people?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(61) Do you try to draw attention to yourself by the way you dress or look?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(62) Do you often make a point of being dramatic and colorful?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(63) Have you often changed your mind about things depending on the people you're with or what you have just read or seen on TV?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No
(64) Do you often express yourself using generalities and very little detail?	1 ☐ Yes → 2 ☐ No - *Go to Section 11A, page 116*	1 ☐ Yes 2 ☐ No } *Go to Section 11A,* Page 116

Section 11A - BEHAVIOR

Statement O **Now I'd like to ask you some questions about experiences you may have had. As I read each experience, please tell me if it has ever happened.**

	1a. In your ENTIRE life, did you EVER . . . *(Repeat entire phrase frequently)*		**b. Did this happen BEFORE you were 15?**	**c. Has this happened SINCE you were 15?**
(1)	**Often cut class, not go to class or go to school and then leave without permission?**	1 ☐ Yes → 2 ☐ No - *Go to next experience*	Ask Before 13 1 ☐ Yes 2 ☐ No	Ask Since 13 1 ☐ Yes } *Go to next* 2 ☐ No } *experience*
(2)	**Stay out late at night even though your parents told you to stay home?**	1 ☐ Yes → 2 ☐ No - *Go to next experience*	Ask Before 13 1 ☐ Yes 2 ☐ No	Ask Since 13 1 ☐ Yes } *Go to next* 2 ☐ No } *experience*
(3)	**Have a time when you bullied or pushed people around or tried to make them afraid of you?**	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes } *Go to next* 2 ☐ No } *experience*
(4)	**Run away from home overnight at least twice when you were living at home or run away and stay away for a longer time?**	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes } *Go to next* 2 ☐ No } *experience*
(5)	**Have a time when you were absent from work or school a lot, other than the times you were sick or taking care of someone else who was sick?**	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes } *Go to next* 2 ☐ No } *experience*
(6)	**More than once quit a job without knowing where you would find another one?**	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes } *Go to next* 2 ☐ No } *experience*
(7)	**More than once quit a school program without knowing what you would do next?**	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes } *Go to next* 2 ☐ No } *experience*
(8)	**Travel around from place to place for a month or more without making any plans ahead of time or not knowing how long you would be gone or where you were going to work?**	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes } *Go to next* 2 ☐ No } *experience*
(9)	**Have a time that lasted at least 1 month when you had no regular place to live – like living on the street or in a car?**	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes } *Go to next* 2 ☐ No } *experience*
(10)	**Have a time that lasted at least 1 month when you lived with friends, acquaintances or relatives because you didn't really have your own place to live?**	1 ☐ Yes → 2 ☐ No - *Go to next experience, page 117*	1 ☐ Yes 2 ☐ No	1 ☐ Yes } *Go to next* 2 ☐ No } *experience, page 117*

Section 11A - BEHAVIOR (Continued)

1a. Did you EVER . . . *(Repeat entire phrase frequently)*		b. Did this happen BEFORE you were 15?	c. Has this happened SINCE you were 15?
(11) Have a time in your life when you lied a lot, not counting any times you lied to keep from being hurt?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience*
(12) Use a false or made-up name or alias?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience*
(13) Scam or con someone for money, to avoid responsibility or just for fun?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience*
(14) Do things that could have easily hurt you or someone else - like speeding or driving after having too much to drink?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience*
(15) Get more than 3 traffic tickets for reckless or careless driving, speeding, or causing an accident?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience*
(16) Have your driver's license suspended or revoked for moving violations?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience*
(17) Destroy, break, or vandalize someone else's property - like their car, home, or other personal belongings?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience*
(18) Start a fire on purpose to destroy someone else's property or just to see it burn?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience*
(19) Fail to pay off your debts - like moving to avoid paying rent, not making payments on a loan or mortgage, failing to make alimony or child support payments or filing for bankruptcy?	1 ☐ Yes 2 ☐ No } *Go to next experience*		
(20) Steal anything from someone or someplace when no one was around?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience*
(21) Forge someone else's signature - like on a legal document or on a check?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience*
(22) Shoplift?	1 ☐ Yes → 2 ☐ No - *Go to next experience, page 118*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience, page 118*

Section 11A - BEHAVIOR (Continued)

1a. Did you EVER . . . *(Repeat entire phrase frequently)*		**b. Did this happen BEFORE you were 15?**	**c. Has this happened SINCE you were 15?**
(23) Rob or mug someone or snatch a purse?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience*
(24) Make money illegally - like selling stolen property or selling drugs?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience*
(25) Do anything that you could have been arrested for, regardless of whether or not you were caught?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience*
(26) Force someone to have sex with you against their will?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience*
(27) Get into a lot of fights that you started?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience*
(28) Get into a fight that came to swapping blows with someone like a husband, wife, girlfriend or boyfriend?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience*
(29) Use a weapon like a stick, knife, or gun in a fight?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience*
(30) Hit someone so hard that you injured them or they had to see a doctor?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience*
(31) Harass, threaten or blackmail someone?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience*
(32) Physically hurt another person in any other way on purpose?	1 ☐ Yes → 2 ☐ No - *Go to next experience*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to next experience*
(33) Hurt or be cruel to an animal or pet on purpose?	1 ☐ Yes → 2 ☐ No - *Go to Check Item 11.0*	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No } *Go to Check Item 11.0*

CHECK ITEM 11.0	Are at least 3 items marked "Yes" in column a, pages 116 - 118?	1 ☐ Yes 2 ☐ No - *SKIP to Section 11B, page 121*
1d.	**About how old were you the FIRST time SOME of these experiences BEGAN to happen?**	_____ Age
CHECK ITEM 11.1	Are at least 3 items marked "Yes" in 1, column b, pages 116 - 118? Did respondent demonstrate at least 3 behaviors BEFORE age 15?	1 ☐ Yes 2 ☐ No - *SKIP to Check Item 11.2, page 119*

Section 11A - BEHAVIOR (Continued)	
2. **You just mentioned some experiences you had BEFORE you were 15 years old.** **Did any of these experiences you had BEFORE you were 15 years old cause any problems with your family or friends, at school or with the law?**	1 ☐ Yes 2 ☐ No
3. **Did at least 1 of these experiences you mentioned happen BEFORE you were 10 years old?**	1 ☐ Yes 2 ☐ No
3a. **Did at least 3 of these experiences you had BEFORE you were 15 years old happen around the same time or within a 1-year period?**	1 ☐ Yes 2 ☐ No
CHECK ITEM 11.1A *Refer to Check Item 2.0, Section 2A, page 9* Is the respondent a lifetime abstainer of alcohol?	1 ☐ Yes - *SKIP to 5a* 2 ☐ No
4a. **Now I'd like you to think about ALL of the experiences you just mentioned that happened BEFORE you were 15 years old.** **Did ANY of these experiences you had BEFORE you were 15 happen WHILE you were drinking heavily, or AFTER you had been drinking heavily?**	1 ☐ Yes 2 ☐ No - *SKIP to 5a*
b. **Did ALL of these experiences ONLY happen WHILE you were drinking heavily, or AFTER you had been drinking heavily?**	1 ☐ Yes 2 ☐ No
5a. **Did ANY of these experiences you had BEFORE you were 15 happen WHILE you were using or AFTER you had used any medicines or drugs?**	1 ☐ Yes 2 ☐ No - *SKIP to Check Item 11.1B*
b. **Did ALL of these experiences ONLY happen WHILE you were using or AFTER you had used any medicines or drugs?**	1 ☐ Yes 2 ☐ No
CHECK ITEM 11.1B Is "Yes" marked in Check Item 5.3, Section 5, page 77? Did respondent ever have a period of high mood?	1 ☐ Yes 2 ☐ No - *SKIP to Check Item 11.2*
5c. **Did ANY of these experiences you had BEFORE you were 15 happen during a period when you felt extremely excited, elated or hyper or extremely irritable or easily annoyed?**	1 ☐ Yes 2 ☐ No - *SKIP to Check Item 11.2*
d. **Did ALL of those experiences ONLY happen during periods when you felt extremely excited, elated or hyper or extremely irritable or easily annoyed?**	1 ☐ Yes 2 ☐ No
CHECK ITEM 11.2 Are at least 3 items marked "Yes" in 1, column c, or "No" in 1, column b, or "Yes" in 1(19), column a, pages 116 - 118? Did respondent demonstrate at least 3 behaviors SINCE age 15?	1 ☐ Yes 2 ☐ No - *SKIP to Section 11B, page 121*
CHECK ITEM 11.2A *Refer to Check Item 2.0, Section 2A, page 9.* Is the respondent a lifetime abstainer of alcohol?	1 ☐ Yes - *SKIP to 7a* 2 ☐ No
6a. **You mentioned some experiences you had SINCE you were 15 years old.** **Did ANY of these experiences you had SINCE you were 15 happen WHILE you were drinking heavily, or AFTER you had been drinking heavily?**	1 ☐ Yes 2 ☐ No - *SKIP to 7a*
b. **Did ALL of these experiences ONLY happen WHILE you were drinking heavily, or AFTER you had been drinking heavily?**	1 ☐ Yes 2 ☐ No
7a. **Did ANY of these experiences you had SINCE you were 15 happen WHILE you were using or AFTER you had used any medicines or drugs?**	1 ☐ Yes 2 ☐ No - *SKIP to Check Item 11.2B, page 120*

Section 11A - BEHAVIOR (Continued)	
7b. **Did ALL of these experiences ONLY happen WHILE you were using or AFTER you had used medicine or drugs?**	1 ☐ Yes 2 ☐ No
CHECK ITEM 11.2B Is "Yes" marked in Check Item 5.3, Section 5, page 77? Did respondent ever have a period of high mood?	1 ☐ Yes 2 ☐ No - *SKIP to Check Item 11.3*
7c. **Did ANY of the experiences you had SINCE you were 15, happen during a time when you felt extremely excited, elated or hyper or extremely irritable or easily annoyed?**	1 ☐ Yes 2 ☐ No - *SKIP to Check Item 11.3*
d. **Did ALL of those experiences ONLY happen during periods when you felt extremely excited, elated or hyper or extremely irritable or easily annoyed?**	1 ☐ Yes 2 ☐ No
CHECK ITEM 11.3 Is at least 1 item marked "Yes" in 1(17) - 1(33), column c, or "No" in 1(17) - 1(33), column b, or "Yes" in 1(19), column a, pages 117 - 118? Has respondent ever destroyed or stolen property or mistreated or harmed another person?	1 ☐ Yes 2 ☐ No - *SKIP to Section 11B, page 121*
8. **You mentioned some experiences that you've had in your life when you (destroyed property/stole something/ mistreated or harmed another person).** **(a)** **Since (this/these things) happened, have you regretted doing (this/these things) or wished (it/they) had never happened?**	1 ☐ Yes 2 ☐ No
(b) **Did you feel you had a right to do (this/these things) or feel that the other people deserved what they got?**	1 ☐ Yes 2 ☐ No } *Go to Section 11B, page 121*

Section 11B - FAMILY HISTORY - IV

Now I would like to ask you about whether any of your relatives, regardless of whether or not they are now living, have ever had behavior problems.

(SHOW FLASHCARD 26)

Statement P

By behavior problems I mean being cruel to people or animals, fighting or destroying property, trouble keeping a job or paying bills, being impulsive, reckless or not planning ahead, lying or conning people or getting arrested. These people also do not seem to care if they hurt others and often have problems at an early age such as truancy, staying out all night or running away.

(REFER TO FLASHCARD FREQUENTLY)

1. **In your judgement, did your blood or natural father have some of these behavior problems like this ANY time in his life?**	1 ☐ Yes 2 ☐ No 99 ☐ DK
2. **Did your blood or natural mother have some of these behavior problems like this ANY time is her life?**	1 ☐ Yes 2 ☐ No 99 ☐ DK
3. **(Did your full brother have/How many of your full brothers had) some of these behavior problems at ANY time in (his life/their lives)?**	1 ☐ Yes 2 ☐ No OR _____ Number 0 ☐ None
4. **(Did your full sister have/How many of your full sisters had) some of these behavior problems at ANY time in (her life/ their lives)?**	1 ☐ Yes 2 ☐ No OR _____ Number 0 ☐ None
5. **(Did your natural son have/How many of your natural sons had) some of these behavior problems at ANY time in (his life/their lives)?**	1 ☐ Yes 2 ☐ No OR _____ Number 0 ☐ None
6. **(Did your natural daughter have/How many of your natural daughters had) some of these behavior problems at ANY time in (her life/their lives)?**	1 ☐ Yes 2 ☐ No OR _____ Number 0 ☐ None
7. **(Did your natural father's full brother have/How many of your natural father's full brothers had) some of these behavior problems at ANY time in (his life/their lives)?**	1 ☐ Yes 2 ☐ No OR _____ Number 0 ☐ None
8. **(Did your natural father's full sister have/How many of your natural father's full sisters had) some of these behavior problems at ANY time in (her life/their lives)?**	1 ☐ Yes 2 ☐ No OR _____ Number 0 ☐ None
9. **(Did your natural mother's full brother have/How many of your natural mother's full brothers had) some of these behavior problems at ANY time in (his life/ their lives)?**	1 ☐ Yes 2 ☐ No OR _____ Number 0 ☐ None
10. **(Did your natural mother's full sister have/How many of your natural mother's full sisters had) some of these behavior problems at ANY time in (her life/their lives)?**	1 ☐ Yes 2 ☐ No OR _____ Number 0 ☐ None

Section 11B - FAMILY HISTORY - IV (Continued)	
11. **Did your natural grandfather on your father's side have some of these behavior problems at ANY time in his life?**	1 ☐ Yes 2 ☐ No 99 ☐ DK
12. **Did your natural grandmother on your father's side have some of these behavior problems at ANY time in her life?**	1 ☐ Yes 2 ☐ No 99 ☐ DK
13. **Did your natural grandfather on your mother's side have some of these behavior problems at ANY time in his life?**	1 ☐ Yes 2 ☐ No 99 ☐ DK
14. **Did your natural grandmother on your mother's side have some of these behavior problems at ANY time in her life?**	1 ☐ Yes 2 ☐ No 99 ☐ DK *Go to Section 12, page 123*

Appendix E

Acronyms and abbreviations

- DSM: Diagnostic and Statistical Manual of Mental Disorders.
- NIAAA: National Institute on Alcohol Abuse and Alcoholism.
- NESARC: National Epidemiologic Survey on Alcohol and Related Conditions.
- BNP: Bayesian nonparametric.
- IBP: Indian Buffet Process.
- DP: Dirichlet Process.
- HDP: Hierarchical Dirichlet process.
- HMM: Hidden Markov model.
- CRP: Chinese Restaurant Process.
- MCMC: Markov Chain Monte Carlo.
- EP: Expectation Propagation.
- MH: Metropolis-Hastings.
- MAP: *Maximum a Posteriori*.
- PD: Personality disorders.
- BMI: Body mass index.

Appendix F

Notation

- N: Number of objects or observations.
- D: Dimensionality of the observations.
- $\mathbf{X}$: $N \times D$ observation matrix.
- $\mathbf{x}_n$: n-th row vector of matrix $\mathbf{X}$.
- $\mathbf{x}^d$: d-th column vector of matrix $\mathbf{X}$.
- x_n^d: each element in matrix $\mathbf{X}$.
- R: number f categories in the observation matrix $\mathbf{X}$.
- K: Number of latent variables.
- K_+: Number of active latent variables.
- $\mathbf{Z}$: $N \times K$ binary latent feature matrix.
- $\mathbf{z}_n$: n-th row vector of matrix $\mathbf{Z}$.
- z_{nk}: each element in matrix $\mathbf{Z}$.
- α: Concentration parameter of the IBP.
- $\mathbf{W}$: $N \times K$ severity matrix.
- $\mathbf{w}_n$: n-th row vector of matrix $\mathbf{W}$.
- w_{nk}: each element in matrix $\mathbf{W}$.

- $\mathbf{B}^d$: $K \times R$ weighting matrix associated to dimension d of the observation matrix $\mathbf{X}$.
- $\mathbf{b}^d_{\cdot r}$: r-th column vector of $\mathbf{B}^d$.
- b^d_{kr}: each element in matrix $\mathbf{B}^d$.
- $\mathbf{b}^d_0$: K-length bias vector associated to dimension d of the observation matrix $\mathbf{X}$.
- b^d_{0r}: each element in $\mathbf{b}^d_0$.
- $\mathbf{Y}^d$: $N \times R$ matrix that contains the Gaussian auxiliary variable y^d_n needed for the IBP model for heterogeneous databases.
- y^d_n: Gaussian auxiliary variable needed for the IBP model for heterogeneous databases.
- θ^d_r: Gaussian thresholds that divide the real line into the number of categories for ordinal observations in the IBP model for heterogeneous databases.
- Ψ^d: Set of auxiliary variables needed to obtain the observations $\mathbf{x}^d$ given $\mathbf{Y}^d$ in the IBP model for Heterogeneous databases.
- $p(\cdot)$: probability distribution function of a random variable.
- $p(x|y)$: conditional pdf of x given y.
- $x \sim p(x)$: The random variable x is distributed as $p(x)$.
- $\mathcal{N}(x|\mu, \sigma^2_x)$: Normal distribution with variable x, mean μ and variance σ^2_x.
- σ^2_x: Variance of variable x.
- $\Phi(\cdot)$: Cumulative density function of the standard normal distribution.
- $\mathbb{E}_{p(x)}[\cdot]$: Expectation with respect to the distribution $p(x)$.
- $f(x)$: Function of x.
- $f^{-1}(\cdot)$: Inverse function of function $f(\cdot)$.
- ∇f: Gradient of function f.
- $\nabla\nabla f$: Hessian of function f.
- $\delta(\cdot)$: Kronecker delta function.

Bibliography

[1] Summary of national strategy for suicide prevention: Goals and objectives for action, 2007. Available at: http://www.sprc.org/library/nssp.pdf.

[2] Aleman A. and D. Denys. Mental health: A road map for suicide research and prevention. *Nature International Weekly of Science - Comment*, 509, 2014.

[3] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications, New York, 1972.

[4] D. Aldous. Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, pages 1–198. Springer, Berlin, 1985.

[5] R. An and H. Zhu. U.s. out-of-pocket health care expenses for mental disorders, 1996-2011. *Psychiatric Services*, 65, 2014.

[6] D. Blackwell and J. B. Macqueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1:353–355, 1973.

[7] C. Blanco, R. F. Krueger, D. S. Hasin, S. M. Liu, S. Wang, B. T. Kerridge, T. Saha, and M. Olfson. Mapping common psychiatric disorders: Structure and predictive validity in the National Epidemiologic Survey on Alcohol and Related Conditions. *Journal of the American Medical Association Psychiatry*, 70(2):199–208, 2013.

[8] D. M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, April 2012.

[9] D. M. Blei, T. L. Griffiths, Jordan M. I., and J. B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In

Advances in Neural Information Processing Systems, page 2003. MIT Press, 2004.

[10] D. M. Blei and M. I. Jordan. Variational inference for dirichlet process mixtures. *Bayesian Anal.*, 1(1):121–143, 03 2006.

[11] D. M. Blei and J. D. Lafferty. A correlated topic model of Science. *Annals of Applied Statistics*, 1(1):17–35, August 2007.

[12] G. Bouchard. Efficient bounds for the softmax and applications to approximate inference in hybrid models. *Advances in Neural Information Processing Systems), Workshop on Approximate Inference in Hybrid Models*, 2007.

[13] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004.

[14] G. K. Brown, T. Ten Have, G. R. Henriques, S.X. Xie, J.E. Hollander, and A. T. Beck. Cognitive therapy for the prevention of suicide attempts: a randomized controlled trial. *Journal of the American Medical Association*, 294(5):563–570, 2005.

[15] W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *J. Mach. Learn. Res.*, 6:1019–1041, December 2005.

[16] B. Cseke and T. Heskes. Approximate marginals in latent Gaussian models. *J. Mach. Learn. Res.*, 12:417–454, February 2011.

[17] B. P. Dohrenwend. Sociocultural and social-psychological factors in the genesis of mental disorders. *Journal of Health and Social Behavior*, 16(4):365–392, 1975.

[18] F. Doshi-Velez and Z. Ghahramani. Accelerated sampling for the indian buffet process. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 273–280, New York, NY, USA, 2009. ACM.

[19] F. Doshi-Velez, K. T. Miller, J. Van Gael, and Y. W. Teh. Variational inference for the Indian buffet process, 2009.

[20] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1994.

[21] T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[22] E. Fox, E. Sudderth, M. I. Jordan, and A. Willsky. A sticky HDP-HMM with application to speaker diarization. *Annals of Applied Statistics*, 5(2A), 2011.

[23] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. A sticky HDP-HMM with application to speaker diarization. *Annals of Applied Statistics*, 5(2A):1020–1056, 2011.

[24] A. Freeman, J. Pretzer, B. Fleming, and K. M. Simon. Avoidant and dependent personality disorders. In *Clinical Applications of Cognitive Therapy*, pages 267–290. Springer US, 1990.

[25] M. Girolami and S. Rogers. Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Computation*, 18:2006, 2005.

[26] M. Girolami and M. Zhong. Data integration for classification problems employing Gaussian process priors. *Advances in Neural Information Processing Systems*, 19:465–472, 2007.

[27] T. L. Griffiths and Z. Ghahramani. The Indian buffet process: an introduction and review. *Journal of Machine Learning Research*, 12:1185–1224, 2011.

[28] E. C. Harris and B. Barraclough. Suicide as an outcome for mental disorders. a meta-analysis. *The British Journal of Psychiatry*, 170:205–228, 1997.

[29] D. A. Harville. *Matrix Algebra From a Statistician's Perspective*. Springer-Verlag, 1997.

[30] S. Haykin. *Adaptive Filter Theory*. Prentice Hall, 2002.

[31] A. B. Hollingshead and F. C. Redlich. Social stratification and psychiatric disorders. *American Sociological Review*, 18(2):163–169, 1953.

[32] T. Insel, B. Cuthbert, M. Garvey, R. Heinssen, D. S. Pine, K. Quinn, C. Sanislow, and P. Wang. Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, 167(7):748–751, 2010.

[33] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[34] M. I. Jordan. *Hierarchical models, nested models and completely random measures*. Springer, New York, (NY), 2010.

[35] M. I. Jordan. What are the open problems in Bayesian statistics? *ISBA Bulletin*, 18(1):1–4, 2011.

[36] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, November 1999.

[37] K. S. Kendler, C. A. Prescott, J. Myers, and M. C. Neale. The structure of genetic and environmental risk factors for common psychiatric and substance use disorders in men and women. *Archives of General Psychiatry*, 60(9):929–937, Sep 2003.

[38] R. C. Kessler, P. Berglund, G. Borges, M. Nock, and P. S. Wang. Trends in suicide ideation, plans, gestures, and attempts in the united states, 1990-1992 to 2001-2003. *Journal of the American Medical Association*, 293(20):2487–2495, 2005.

[39] R. C. Kessler, K. A. McGonagle, M. Swartz, D. G. Blazer, and C. B. Nelson. Sex and depression in the national comorbidity survey i: Lifetime prevalence, chronicity and recurrence. *Journal of Affective Disorders*, 29(2):85–96, 1993.

[40] D. A. Knowles and Z. Ghahramani. Nonparametric Bayesian sparse factor models with application to gene expression modelling. *Annals of Applied Statistics*, 5(2B):1534–1552, 2011.

[41] D. A. Knowles and Z. Ghahramani. Nonparametric Bayesian sparse factor models with application to gene expression modelling. *Annals of Applied Statistics*, 5(2B):1534–1552, 2011.

[42] R. Kotov, C. J. Ruggero, R. F. Krueger, D. Watson, Q. Yuan, and M. Zimmerman. New dimensions in the quantitative classification of mental illness. *Archives of General Psychiatry*, 68(10):1003–1011, 2011.

[43] R. F. Krueger. The structure of common mental disorders. *Archives of General Psychiatry*, 56(10):921–926, 1999.

[44] K. Krysinska and G. Martin. The struggle to prevent and evaluate: application of population attributable risk and preventive fraction to suicide prevention research. *Suicide and Life-Threatening Behavior*, 39(5):548–557, 2009.

[45] M. Kuss and C. E. Rasmussen. Assessing approximate inference for binary Gaussian process classification. *Journal of Machine learning research*, 6:1679–1704, Oct. 2005.

[46] X.-B. Li. A Bayesian approach for estimating and replacing missing categorical data. *J. Data and Information Quality*, 1(1):3:1–3:11, June 2009.

[47] A. Y. Lo. On a class of bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics*, 12(1):351–357, 03 1984.

[48] D. J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA, 2002.

[49] J. J. Mann, A. Apter, J. Bertolote, A. Beautrais, D. Currier, A. Haas, U. Hegerl, J. Lonnqvist, K. Malone, A. Marusic, L. Mehlum, G. Patton, M. Phillips, W. Rutz, Z. Rihmer, A. Schmidtke, D. Shaffer, M. Silverman, Y. Takahashi, A. Varnik, D. Wasserman, P. Yip, and H. Hendin. Suicide prevention strategies: a systematic review. *The Journal of the American Medical Association*, 294(16):2064–2074, 2005.

[50] T. L. Mark, K. R. Levit, R. Vandivort-Warren, J. A. Buck, and R. M. Coffey. Changes in us spending on mental health and substance abuse treatment, 19862005, and implications for policy. *Health Affairs*, 30(2):284–292, 2011.

[51] T. P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, pages 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[52] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[53] N. O'Mahony, B. Florentino-Liano, J. J. Carballo, E. Baca-Garcia, and A. Artes Rodriguez. Objective diagnosis of ADHD using IMUs. *Medical Engineering and Physics*, 36(7):922 – 926, 2014.

[54] M. Opper and O. Winther. Expectation consistent approximate inference. *J. Mach. Learn. Res.*, 6:2177–2204, December 2005.

[55] M. Oquendo, E. Baca-Garcia, A. Artés-Rodríguez, F. Perez-Cruz, H. Galfalvy, H. Blasco-Fontecilla, D. Madigan, and N. Duan. Machine learning and data mining: strategies for hypothesis generation. *Molecular psychiatry*, 17(10):956–959, 2012.

[56] M. A. Oquendo, E. B. García, J. J. Mann, and J. Giner. Issues for DSM-V: suicidal behavior as a separate diagnosis on a separate axis. *The American Journal of Psychiatry*, 165(11):1383–1384, November 2008.

[57] P. Orbanz and Y. W. Teh. Modern bayesian nonparametrics. In *Tutorial at Neural Information Processing Systems*, http://www.youtube.com/watch?v=F0_ih7THV94, december 2011.

[58] J. W. Paisley, D. M. Blei, and M. I. Jordan. Variational Bayesian inference with stochastic search. In *International Conference on Machine Learning 29. (ICML'12)*. icml.cc / Omnipress, 2012.

[59] W. Parry. DSM-5 controversy: What is normal and what is not? In *HuffPost*, 2013.

[60] L. S. Penrose. The importance of statistics in psychiatry. *Journal of the Royal Society of Medicine*, 40(14):863–870, 1947.

[61] Y. Qi, T. P. Minka, R. W. Picard, and Z. Ghahramani. Predictive automatic relevance determination by expectation propagation. In *Proceedings of the twenty-first International conference on Machine Learning*, ICML '04, pages 671–678, 2004.

[62] V. Rao and Y. W. Teh. Fast MCMC sampling for Markov jump processes and continuous time Bayesian networks. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, 2011.

[63] C. E. Rasmussen. The infinite gaussian mixture model. In *In Advances in Neural Information Processing Systems 12*, pages 554–560. MIT Press, 2000.

[64] J. Riihimäki, P. Jylänki, and A. Vehtari. Nested expectation propagation for Gaussian process classification with a multinomial

probit likelihood. *Journal of Machine Learning Research*, 14:75–109, 2013.

[65] C. P. Robert. Simulation of truncated normal variables. *Statistics and computing*, 5(2):121–125, 1995.

[66] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, 2007.

[67] R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using Markov Chain Monte Carlo. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 880–887, New York, NY, USA, 2008. ACM.

[68] E. Salazar, M. Cain, E. Darling, S. Mitroff, and L. Carin. Inferring latent structure from mixed real and categorical relational data. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, pages 1039–1046, New York, NY, USA, July 2012. Omnipress.

[69] R. Savage, K. Heller, Y. Xu, Z. Ghahramani, W. Truman, M. Grant, K. Denby, and D. Wild. R/bhc: fast bayesian hierarchical clustering for microarray data. *BMC Bioinformatics*, 10(1):242, 2009.

[70] ScienceDaily. Big data, for better or worse: 90% of world's data generated over last two years.

[71] M. W. Seeger. Expectation propagation for exponential families. Technical report, 2005.

[72] M. W. Seeger. Bayesian inference and optimal design for the sparse linear model. *J. Mach. Learn. Res.*, 9:759–813, June 2008.

[73] M. W. Seeger and M. I. Jordan. Sparse Gaussian process classification with multiple classes. Technical report, University of California, Berkeley, CA, 2004.

[74] E. Serrano-Drozdowskyj, J. Lopez-Castroman, J.M. Leiva-Murillo, H. Blasco-Fontecilla, R. Garcia-ieto, A. Artes-Rodriguez, C. Morant-Ginestar, C. Blanco, P. Courtet, and E. Baca-Garcia. 1533 a naturalistic study of the diagnostic evolution of schizophrenia. *European Psychiatry*, 28, Supplement 1(0):1 –, 2013.

[75] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[76] J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[77] K. Smith. Trillion-dollar brain drain. *Nature*, 478(7367):15, 2011.

[78] A. Soni. The five most costly conditions, 1996 and 2006: Estimates for the u.s. civilian noninstitutionalized population. *Statistical Brief*, 248(453), 2009.

[79] E. Sudderth and M. Jordan. Shared segmentation of natural scenes using dependent pitman-yor processes. In *Neural Information Processing Systems 21*. MIT Press, Cambridge, MA, 2009.

[80] K. Suominen, M. Henriksson, J. Suokas, E. Isomets, A. Ostamo, and J. Lnnqvist. Mental disorders and comorbidity in attempted suicide. *Acta Psychiatrica Scandinavica*, 94(4):234–240, 1996.

[81] K. Szanto, S. Kalmar, H. Hendin, Z. Rihmer, and J. J. Mann. A suicide prevention program in a region with a very high suicide rate. *Archives of General Psychiatry*, 64(8):914–920, 2007.

[82] Y. W. Teh, D. Görür, and Z. Ghahramani. Stick-breaking construction for the Indian buffet process. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 11, 2007.

[83] M. Titsias. The infinite gamma-Poisson feature model. *Advances in Neural Information Processing Systems*, 19, 2007.

[84] A. Todeschini, F. Caron, and M. Chavent. Probabilistic low-rank matrix completion with adaptive spectral regularization algorithms. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 845–853. Curran Associates, Inc., Dec. 2013.

[85] J. M. Valderas, B. Starfield, B. Sibbald, C. Salisbury, and M. Roland. Defining comorbidity: implications for understanding health and health services. *Annals of family medicine*, 7(4):357–363, 2009.

[86] I. Valera and Z. Ghahramani. General table completion using a bayesian nonparametric model. *Advances in Neural Information Processing Systems*, 2014.

[87] J. Van Gael, Y. W. Teh, and Z. Ghahramani. The infinite factorial hidden Markov model. In *Advances in Neural Information Processing Systems*, volume 21, 2009.

[88] W. A. Vollebergh, J. Ledema, R.V. Bijl, R. de Graaf, F. Smit, and J. Ormel. The structure and stability of common mental disorders: the NEMESIS study. *Archives of General Psychiatry*, 58(6):597–603, Jun 2001.

[89] M.J. Wainwright and M.I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[90] C. Wang and D. M. Blei. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(1):1005–1031, 2013.

[91] S. Weich and G. Lewis. Poverty, unemployment, and common mental disorders: population based cohort study. *BMJ*, 317(7151):115–119, 1998.

[92] M. M. Weissman, Bland R., P. R. Joyce, S. Newman, J.E. Wells, and H.-U. Wittchen. Sex differences in rates of depression: cross-national perspectives. *Journal of Affective Disorders*, 29(23):77 – 84, 1993. Special Issue Toward a New Psychobiology of Depression in Women.

[93] C. K. I. Williams and D. Barber. Bayesian classification with Gaussian Processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:1342–1351, 1998.

[94] S. Williamson, A. Dubey, and E. P. Xing. Parallel Markov chain Monte Carlo for nonparametric mixture models. In *International Conference on Machine Learning (ICML)*, volume 28 of *JMLR Proceedings*, pages 98–106, 2013.

[95] S. Williamson, C. Wang, K. Heller, and D. Blei. The IBP compound Dirichlet process and its application to focused topic modeling. *Proceedings of the 27th Annual International Conference on Machine Learning*, 2010.

[96] J. L. Wolff, B. Starfield, and G. Anderson. Prevalence, expenditures, and complications of multiple chronic conditions in the elderly. *Archives of Internal Medicine*, 162(20):2269–2276, 2002.

[97] M. A. Woodbury. The stability of out-input matrices. *Mathematical Reviews*, 1949.